THE K-SAMPLE PROBLEM WHEN K IS LARGE AND N SMALL

A Dissertation

by

DONGLING ZHAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Statistics

THE K-SAMPLE PROBLEM WHEN K IS LARGE AND N SMALL

A Dissertation

by

DONGLING ZHAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Jeffrey D. Hart
Committee Members,	Jianhua Huang
	Michael Sherman
	Ximing Wu
Head of Department,	Simon J. Sheather

May 2012

Major Subject: Statistics

ABSTRACT

The k-Sample Problem When k is Large and n Small.

(May 2012)

Dongling Zhan, B.S., Fudan University;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Jeffrey D. Hart

The k-sample problem, i.e., testing whether two or more data sets come from the same population, is a classic one in statistics. Instead of having a small number of k groups of samples, this dissertation works on a large number of p groups of samples, where within each group, the sample size, n, is a fixed, small number. We call this as a "Large p, but Small n" setting. The primary goal of the research is to provide a test statistic based on kernel density estimation (KDE) that has an asymptotic normal distribution when p goes to infinity with n fixed.

In this dissertation, we propose a test statistic called $T_p^{(S)}$ and its standardized version, $T^{(S)}$. By using $T^{(S)}$, we conduct our test based on the critical values of the standard normal distribution. Theoretically, we show that our test is invariant to a location and scale transformation of the data. We also find conditions under which our test is consistent. Simulation studies show that our test has good power against a variety of alternatives. The real data analyses show that our test finds differences between gene distributions that are not due simply to location.

To My Family

ACKNOWLEDGEMENTS

It is my great pleasure to thank those who have made this dissertation possible. First of all, my deepest gratitude is owed to my advisor, Dr. Jeffrey D. Hart, for his excellent academic guidance. Dr. Hart has always been there with his knowledge and patience to answer questions regarding my research. He has also taught me that I cannot be too careful when writing a dissertation. I feel very fortunate to have him as my supervisor. Without his guidance, this dissertation would have not come to fruition. Secondly, I would like to thank all of my committee members, Dr. Jianhua Huang, Dr. Michael Sherman, and Dr. Ximing Wu for their time spent reviewing my dissertation, giving me suggestions, and attending my oral exams. They have always provided prompt responses to my questions and concerns.

Over the past few years, the Department of Statistics has provided me with great support, both financially and spiritually. I owe a special thank you to Dr. Michael Longnecker and Mrs. Julie H. Carroll, who gave me the opportunity of teaching undergraduate statistical courses, which not only benefited my research but also left me with unforgettable memories.

I also want to express my gratitude to Dr. Marcia Ory, who encouraged me to continue pursuing my Ph.D. during difficult times.

Last but not least, I especially want to thank my family for their continuous encouragement and support. I would like to thank my parents for bringing me up with freedom to accomplish my goals and my husband, Xiafeng Li, for always standing by me and providing me a home with two lovely children, Allen & Alison Li. It has been a joy to have them with me on this long journey of pursuing my doctorate.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1.1 The k-Sample Problem

The k-sample problem, i.e., testing whether two or more data sets come from the same population, is a classic one in statistics. Unlike most k-sample problems where k indicates a small, fixed number of distributions, this dissertation focuses on testing equality of a large number of distributions, which can be hundreds or even thousands of distributions. Therefore, instead of using k as in a k-sample problem, we use p to indicate the large number of data sets. Another difference from most k-sample problems is that these p data sets have very small sample sizes, usually less than 10. In other words, the k-sample problem in this research is a large p, but small n problem. Because of the small sample sizes for each data set, we usually cannot use the central limit theorem to justify that the mean of each small data set is normally distributed. This leads us to seek a different way to solve such related problems.

The motivation of this research is from microarray analyses, where one often has a large amount of genes, say, 8000 different genes, in an experiment. However, due to time constraints or budget limitations, one may only have a few experimental subjects. For example, there are only four to five mice available in one such experiment. This kind of experimental design will generate data sets in the way aforementioned. Often, we use p to denote the number of genes and n the number of observations for each gene.

The format and style follow that of *Biometrics*.

1.2 Null and Alternative Hypothesis Test

As a matter of fact, a lot of the current research on a large number of small data sets has been related to multiple hypotheses testing, for example, the simultaneous testing for a treatment effect on hundreds, or even thousands of genes. Dudoit, Shafer, and Boldrick (2003) have discussed different approaches to multiple hypotheses testing in the context of DNA microarray experiments. As an example, we may have the following setting: $X_{ij} = T_{ij} - C_{ij}$, where $T_{i1}, \ldots, T_{in}$ are the normalized treatment expression levels for a given gene i, and $C_{i1}, \ldots, C_{in}$ are the corresponding control levels. Define $\mu_i = E(X_{ij})$, $i = 1, \ldots, p$, $j = 1, \ldots, n$. One may be interested in testing the hypotheses:

$$H_{0i} : \mu_i = 0, \ i = 1, \ldots, p, \tag{1.1}$$

where the null hypothesis is considered as no treatment effect on a given gene i.

However, the primary goal of the research in this dissertation is to test the hypothesis that the distribution of $X_{i1}, \ldots, X_{in}$ is the same for $i = 1, \ldots, p$. Our data are a p by n matrix. The sample size, n, is assumed to be the same for each small data set only for notational simplicity. It is straightforward to extend the proposed methodology to accommodate different sample sizes.

Define X_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, n$, to be the data we have. Suppose that X_{ij}, $j = 1, \ldots, n$, have the distribution F_i, $i = 1, \ldots, p$. We are interested in testing if all the F_i's are the same. Thus, the null hypothesis is

$$H_0 : F_1 = F_2 = \cdots = F_p,$$

with the alternative hypothesis being

$$H_A : F_i \neq F_j,$$

for at least some $i \neq j$.

1.3 Background on Testing Equality of Distributions

In recent years, much new methodology has been developed for testing equality of distributions, especially between two populations. For categorical variables, the most popular test is the chi-square goodness-of-fit test (Plackett (1983)), which may be used to test whether the random variables are from a specific family of distributions. For continuous variables, when only mean differences are of interest, the t-test (Moore, McCabe, and Craig (2007)) is perhaps the most popular two-sample test, which includes the independent and the paired two sample t-test. The two-sample Kolmogorov-Smirnov (K-S) test (Stephens (1974)) is a useful and popular nonparametric method for testing whether two samples are from the same distribution. Other traditional two-sample goodness-of-fit tests based on the empirical distribution function (EDF) include the Cramer-von Mises (CvM) and Anderson-Darling (AD) tests (Anderson and Darling (1954), Anderson (1962), Stephens (1974),Stephens (1986)). Some other tests based on ranks are the Wilcoxon signed-rank test and the Mann-Whitney U-test (Corder and Foreman (2009)), which are used to compare two related or unrelated samples respectively. Recently, a test based on empirical characteristic functions (ECF) has also been proposed (Jimenez-Gamero, Albr-Fernandez, Munoz-Garcia, and Chalco-Cano (2009)). There also exist tests based on kernel density estimation (KDE), as introduced by Rosenblatt (1956). Anderson, Hall, and Titterington (1994), Louani (2000), and Cao and Van Keilegom (2006) considered KDE-based tests for the two-sample problem.

As regards testing equality of multiple distributions, Kruskal and Wallis (1952) developed a non-parametric test based on ranks for testing equality of two or more populations. Kiefer (1959) proposed an extension of the K-S and CvM tests to the k-sample settings. Scholz and Stephens (1987) extended the Anderson-Darling test to the k-sample case, too. The classic one-way ANOVA F test (Iversen and Norpoth (1987)) is often applied to com-

pare multiple means or check homogeneity among groups. Homogeneity of variances tests include Bartlett's test (Snedecor and Cochran (1989)), Hartley's F_{max} test (Hartley (1950)), and Levene's test (Levene (1960)). As for the KDE-based tests for the k-sample problem, Martínez-Camblor, Uña-Álvarez, and Corral (2008), proposed a test for the comparison of k samples based on kernel density estimators. Explicitly, they proposed the following test statistic:

$$AC = \int \min\{f_{n_i}(t), \ldots, f_{n_k}(t)\}dt,$$

where f_{n_i} denotes the KDE pertaining to the i-th sample. The authors suggested that the *AC* test may be more powerful than the EDF-based tests. Later, Martínez-Camblor and Uña-Álvarez (2009) compared the results of *AC* test to those of the traditional EDF-based k-sample tests, and to other tests based on the likelihood ratio introduced in the recent literature. Their simulations suggested that KDE-based tests are the most powerful in the considered situations.

1.4 Uniqueness of the Research

As stated in the previous section, much methodology is applied to testing the equality of distributions, including multiple (k) distributions. However, all these tests provide the asymptotic limiting distribution when sample sizes go to infinity, but the number of distributions, k, is a fixed number. For example, Martínez-Camblor, Uña-Álvarez, and Corral (2008) established the asymptotic normality of the *AC* test statistic when the sample sizes go to infinity. To the best of our knowledge, there is no literature on testing equality of k distributions when the sample sizes are fixed but k goes to infinity. This research is unique in that we have a very small sample size n, which is usually less than 10. We also have p distributions, where p is a very large number, and we derive the limiting distribution of the test statistic when n is fixed, but p goes to infinity.

The rest of the chapters are organized as follows: we propose our methodology based on kernel density estimates in Chapter II. The power of the tests is investigated by simulation in Chapter III. In Chapter IV, we apply our method to a real data set. In Chapter V, a summary of the dissertation is given and future research is discussed. Supplemental materials including proofs, R-codes and other useful information are available in the Appendices.

CHAPTER II

METHODOLOGY

In this chapter, we firstly have an overview of kernel density estimation, including definitions and bandwidth selection issues in Section 2.1 and 2.3. We propose the test statistic based on kernel density estimates (KDE) in Section 2.2. In Section 2.4, we provide an alternative version of the test statistic and obtain its asymptotic distribution in Section 2.5. An invariance property of the standardized test statistic is discussed in Section 2.6. Finally, an asymptotic power analysis is given in Section 2.7.

2.1 A Review of Kernel Density Estimation

In statistics, a variety of approaches to density estimation has been used, both parametric and non-parametric. Kernel density estimation, introduced by Rosenblatt (1956) and Parzen (1962), is the most popular non-parametric way of estimating the probability density function of a random variable. The books by Silverman (1986) and Wand and Jones (1995) discuss nonparametric density estimation based on kernel methods in great detail. This section presents a brief overview of kernel density estimation for those who are not familiar with the subject.

Let $X_1, \ldots, X_n$ be a random sample from a density f. The kernel density estimate of f at the point x is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where h is known as the bandwidth, and $K(\cdot)$ is called a kernel, which is assumed to be a symmetric function satisfying the following conditions, according to Silverman (1986),

p.38:

$$\begin{aligned}\int_{-\infty}^{\infty} K(t)dt &= 1,\\ \int_{-\infty}^{\infty} tK(t)dt &= 0,\\ \int_{-\infty}^{\infty} t^2K(t)dt &= k_2 \neq 0.\end{aligned}$$

There are many choices for $K(\cdot)$. Some of the kernels given in Silverman (1986) are listed as follows, where I is the indicator function defined as follows. For any set S,

$$I_S(x) = \begin{cases} 1 \quad , \quad x \in S, \\ 0 \quad , \quad x \in S. \end{cases}$$

Alternatively, the notation $I(S)$ will be used for I_S.

- Uniform Kernel: $K(t) = U[-1,1] = \frac{1}{2}I(|t| \leq 1)$;
- Triangle Kernel: $K(t) = (1-|t|)I(|t| \leq 1)$;
- Epanechnikov Kernel: $K(t) = \frac{3}{4}(1-t^2)I(|t| \leq 1)$;
- Standard Normal Kernel: $K(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})$.

The shapes of these kernel functions are shown in Figure 1.

The choice of the kernel is not vital as shown by Epanechnikov (1969) and Silverman (1978). Among the many kernels, the standard normal kernel is the most popular one, and hence we adopt it as our kernel function for the purpose of this dissertation.

However, selecting an appropriate bandwidth (h) is much more important than choosing a kernel function. There exists much theory and methodology to estimate an optimal bandwidth. There are two broad classes of methods: classic methods such as cross-validation and Mallows' C_p, and the plug-in methods (see details in Loader (1999)). Since

the 1980's, bandwidth selection has experienced a wide explosion of interest. Sheather (2004) reviewed different methods of choosing a value of h. As a rule of thumb, wider bandwidths result in smoother density estimates. If a very small bandwidth is chosen, an implausibly wiggly density curve is the result. Too big a bandwidth can result in a density curve that is over-smoothed.

Define $N(\mu, \sigma^2)$ to be the normal distribution with mean μ and standard deviation σ. Figure 2 displays examples of different density curve estimates by using different bandwidths, where samples are randomly generated from $N(0, 1)$ with sample sizes 5 or 100. The red solid curves are the standard normal densities and the black dashed curves are the kernel density estimates. The top two plots use a small bandwidth of 0.1, each giving a very poor density estimate. The lower two plots plot the kernel density estimates using the optimal bandwidth. It can be seen that even the optimal bandwidth doesn't give an estimate close to the standard normal when the sample size is as small as 5. This finding motivated our bandwidth selection method in Section 2.3, since our sample size n is assumed to be very small. It suggests that we should pick a bandwidth larger than the optimal bandwidth for our kernel density estimates. However, as sample size increases to 100, we notice that the optimal bandwidth produces a satisfying density estimate that is very close to the true density. In both cases, too small a bandwidth (bandwidth=0.1 in this example) makes the density estimates very wiggly, but a very large bandwidth of 2 makes the estimates over-smoothed.

Due to the fact that different bandwidths result in different kernel estimates, as shown in Figure 2, the bandwidth plays a significant role in kernel density estimation. It is also known that the optimal bandwidth will tend to zero if the sample size tends to infinity.

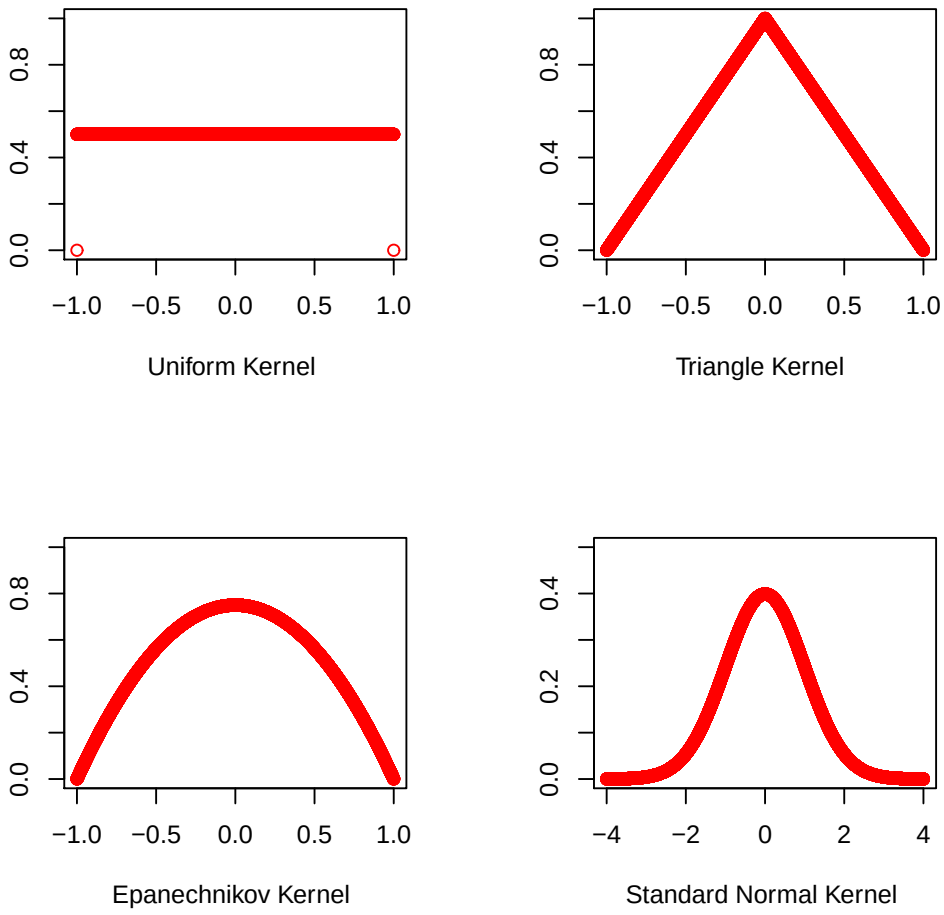

Figure 1: Plots of several types of kernel functions which are commonly used: the uniform, triangle, Epanechnikov and standard normal kernels.

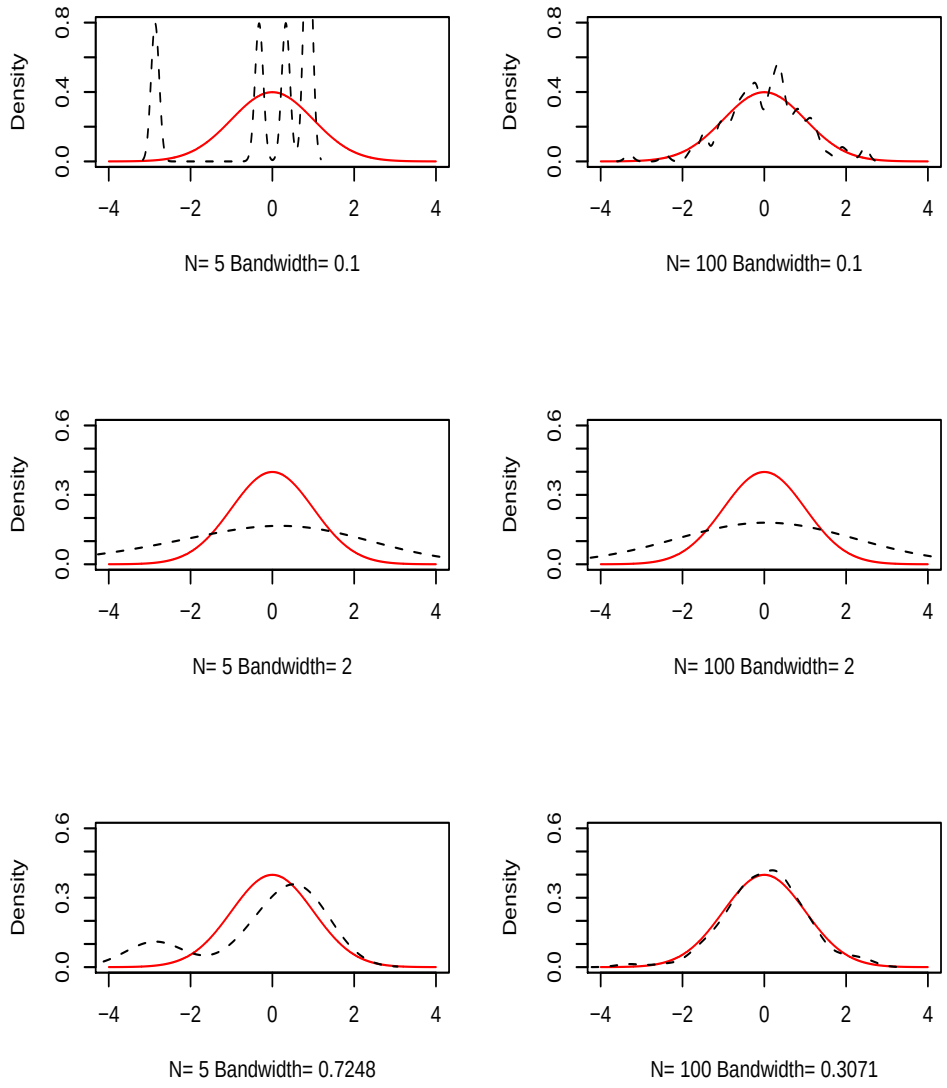

Figure 2: Plots of kernel density estimates (black dashed curves) with sample sizes 5 or 100 and different bandwidths. The last two curves use the optimal bandwidth. The red solid curves are the true densities, which are the standard normal distribution.

2.2 Original Test Statistic: T_p

It is well-established that comparing distributions on the density scale is more powerful than on the cumulative distribution function (CDF) scale (Rayner and Best (1989), Eubank, Hart, and LaRiccia (1993) and Hart (1997) on pp.239-240). Therefore, we intend to propose a test statistic based on kernel density estimation.

For the purpose of notational simplicity, sample sizes n_i, $i = 1, \ldots, p$, are assumed to be the same as n for each small data set. We may extend the proposed methodology to accommodate different sample sizes in the future. At this point, our observed data are $\mathbf{X}$, which is a p by n matrix as follows:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix}.$$

Young and Bowman (1995) proposed a natural test statistic by analogy with one-way analysis of variance, which was used in the case of testing the equality of two or more smooth curves. We use the same principle to propose a test statistic T_p by using kernel density estimation. The test statistic has the following form:

$$T_p = \frac{n}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} \left(\hat{f}_h(x|i) - \hat{f}_h^{(i)}(x) \right)^2 dx, \tag{2.1}$$

where $\hat{f}_h(\cdot|i)$ is a kernel density estimate computed from $X_{i1}, \ldots, X_{in}, i = 1, \ldots, p$, and $\hat{f}_h^{(i)}(x)$ is a kernel density estimate computed from all the data excluding the i-th data set. More explicitly, we have

$$\begin{aligned} \hat{f}_h(x|i) &= \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_{ij}}{h} \right), i = 1, \ldots, p, \\ \hat{f}_h^{(i)}(x) &= \frac{1}{n(p-1)h} \sum_{\{k=1, k \neq i\}}^{p} \sum_{j=1}^{n} K\left(\frac{x - X_{kj}}{h} \right), i = 1, \ldots, p, \end{aligned}$$

where $K(\cdot)$ is the kernel function and h is the bandwidth.

Suppose that $\gamma(x|i)$ is the true density for the i-th small data set, $i = 1, \ldots, p$. We intend to test the null hypothesis

$$H_0 : \gamma(x|1) = \gamma(x|2) = \cdots = \gamma(x|p) = \gamma(x),$$

where γ is unspecified.

As $p \to \infty$, $\hat{f}_h^{(i)}(x)$ is a consistent estimator of $\gamma(x)$ under the null hypothesis. Thus, under the null hypothesis, the average discrepancy between $\hat{f}_h(x|i)$ and $\hat{f}_h^{(i)}(x)$ will be relatively small. In other words, we will reject the null hypothesis if the value of the test statistic T_p is relatively large.

2.3 Bandwidth Selection

As mentioned in Section 2.1, the bandwidth plays an important role in kernel density estimation. Selection of the bandwidth is also one of our tasks while computing the statistic T_p in (2.1). Although the theory to calculate an optimal bandwidth is beyond our discussion, we may utilize the existing methodology to modify our bandwidth. Since the sample size n is bounded and usually is very small, our bandwidth h should not tend to zero as p goes to infinity. It is desirable to find a relatively "large" bandwidth that can be applied to each small data set. Terrell and Scott (1985) and Terrell (1990) developed a bandwidth selection method based on the maximal smoothing principle. This method was proposed to produce an oversmoothed density estimate, which can be considered as a conservative choice for h in our case. According to the maximal smoothing principle, one should use the following bandwidth if using the standard normal kernel (see Sheather (2004)):

$$h_{OS} = 1.144sn^{-1/5},$$

where s is the sample standard deviation.

In our case, we have fixed n. The only remaining question is: how do we obtain s? We suggest using an "average" standard deviation, i.e., our s^2 is chosen to be the mean of all the p sample variances calculated from the data. We call our s "s_{pool}", indicating the standard deviation obtained from pooling information from all data. This is calculated as follows:

$$s^2_{pool} = \frac{1}{N-p}\sum_{i=1}^{p}(n_i - 1)s_i^2,$$

where $N = \sum_{i=1}^{p} n_i$, and s_i^2 is the sample variance from the i-th data set.

Therefore, when $K(\cdot)$ is the standard normal kernel, and $n_i = n$, the modified standard deviation is: $s_{pool} = \sqrt{\frac{1}{p}\sum_{i=1}^{p} s_i^2}$, and our final choice for the bandwidth is:

$$h = 1.144 s_{pool} n^{-1/5}.$$

This choice of h is highly suboptimal for $\hat{f}_h^{(i)}(x)$. However, the advantage of using the same h for $\hat{f}_h^{(i)}(\cdot)$ as for $\hat{f}_h(\cdot|i)$, $i = 1, \ldots, p$, will be discussed later.

2.4 Alternative Test Statistic: $T_p^{(S)}$

It is shown in Appendix I that test statistic T_p has the asymptotic normal distribution under the null. However, closer examination of T_p shows that proper centering does not even require an estimate of ET_p. This examination also reveals an easier way to compute T_p. To compute T_p, one way is based on the specific formula as in (2.1). We firstly compute the two kernel density estimates, $\hat{f}_h(x|i)$ and $\hat{f}_h^{(i)}(x)$, for each i. We then calculate the integral of $\left(\hat{f}_h(x|i) - \hat{f}_h^{(i)}(x)\right)^2$. This usually results in intensive computation. To enhance the speed, we notice that T_p can be rewritten so that "deleting the i-th data set" is avoided.

As a matter of fact, $\hat{f}_h^{(i)}(x)$ can be written in terms of $\hat{f}_h(x|i)$ and $\hat{f}_h(x)$ as follows, where $\hat{f}_h(x)$ is the kernel density estimate from all the pooled data with the same bandwidth

h. We have

$$\begin{aligned}
& \hat{f}_h^{(i)}(x) \\
= & \frac{1}{n(p-1)h}\left(\sum_{\{k=1,k\neq i\}}^{p}\sum_{j=1}^{n} K\left(\frac{x-X_{kj}}{h}\right)+\sum_{j=1}^{n} K\left(\frac{x-X_{ij}}{h}\right)-\sum_{j=1}^{n} K\left(\frac{x-X_{ij}}{h}\right)\right) \\
= & \frac{1}{n(p-1)h}\left(\sum_{k=1}^{p}\sum_{j=1}^{n} K\left(\frac{x-X_{kj}}{h}\right)-\sum_{j=1}^{n} K\left(\frac{x-X_{ij}}{h}\right)\right) \\
= & \frac{p}{p-1}\cdot\frac{1}{nph}\sum_{k=1}^{p}\sum_{j=1}^{n} K\left(\frac{x-X_{kj}}{h}\right)-\frac{1}{p-1}\cdot\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-X_{ij}}{h}\right) \\
= & \frac{p}{p-1}\cdot\hat{f}_h(x)-\frac{1}{p-1}\cdot\hat{f}_h(x|i).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{f}_h(x|i)-\hat{f}_h^{(i)}(x) &= \hat{f}_h(x|i)-\frac{p}{p-1}\cdot\hat{f}_h(x)+\frac{1}{p-1}\cdot\hat{f}_h(x|i) \\
&= \frac{p}{p-1}\left(\hat{f}_h(x|i)-\hat{f}_h(x)\right),
\end{aligned}$$

for $i=1,\ldots,p$, and T_p can be rewritten as follows:

$$\begin{aligned}
T_p &= \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\left(\hat{f}_h(x|i)-\hat{f}_h^{(i)}(x)\right)^2 dx \\
&= \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\frac{p^2}{(p-1)^2}\left(\hat{f}_h(x|i)-\hat{f}_h(x)\right)^2 dx \\
&= \frac{np^2}{(p-1)^2}\frac{1}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\left(\hat{f}_h(x|i)-\hat{f}_h(x)\right)^2 dx.
\end{aligned}$$

Now note that

$$S\equiv\frac{1}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\left(\hat{f}_h(x|i)-\hat{f}_h(x)\right)^2 dx=\frac{1}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\hat{f}_h^2(x|i)dx-\int_{-\infty}^{\infty}\hat{f}_h^2(x)dx,$$

which is true because $\hat{f}_h(x)$ is the mean of $\hat{f}_h(x|1),\ldots,\hat{f}_h(x|p)$.

Since we choose the standard normal density, $\phi(\cdot)$, as our kernel function, we can take advantage of the convolution property between two normal distributions (see Appendix II)

and obtain the following:

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}_h^2(x|i)dx &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{j=1}^{n}\phi\left(\frac{x-X_{ij}}{h}\right)\cdot\frac{1}{nh}\sum_{l=1}^{n}\phi\left(\frac{x-X_{il}}{h}\right)dx \\ &= \frac{1}{n^2h^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\int_{-\infty}^{\infty}\phi\left(\frac{x-X_{ij}}{h}\right)\phi\left(\frac{x-X_{il}}{h}\right)dx.\end{aligned}$$

Using the change of variable: $y = \frac{x-X_{il}}{h}$, the last expression is

$$\begin{aligned}&\frac{h}{n^2h^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\int_{-\infty}^{\infty}\phi\left(y-\frac{X_{ij}-X_{il}}{h}\right)\phi(y)dy \\ =\ &\frac{1}{n^2h}\sum_{j=1}^{n}\sum_{l=1}^{n}\int_{-\infty}^{\infty}\phi\left(\frac{X_{ij}-X_{il}}{h}-y\right)\phi(y)dy \\ =\ &\frac{1}{n^2\sqrt{2}h}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right).\end{aligned}$$

And similarly, we have

$$\int_{-\infty}^{\infty}\hat{f}_h^2(x)dx = \frac{1}{p^2n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right),$$

and therefore, we may write

$$\begin{aligned}S &= \frac{(p-1)}{p}\frac{1}{n\sqrt{2}h}\phi(0) + \frac{(p-1)}{p^2}\frac{1}{n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n}\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right) \\ &\quad -\frac{1}{p^2n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right).\end{aligned}$$

It now becomes obvious how to ensure that the numerator has mean 0. Firstly, drop the term depending on $\phi(0)$ since it contains no information about the underlying densities. Secondly, up to known multipliers, the two sums have the same expectations under H_0. Therefore, we can just modify the multiplier of the second sum. It turns out that the

following quantity has mean 0 under H_0:

$$\begin{aligned}
\widetilde{S} &= \frac{(p-1)}{p^2}\frac{1}{n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n}\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right) \\
&\quad -\frac{(n-1)}{p^2n^3\sqrt{2}h}\sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right) \\
&= \left(\frac{p-1}{p}\right)\left(\frac{n-1}{n}\right)\left[\frac{1}{pn(n-1)\sqrt{2}h}\sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n}\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right)\right. \\
&\quad \left.-\frac{1}{p(p-1)n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right)\right] \\
&= \left(\frac{p-1}{p}\right)\left(\frac{n-1}{n}\right)[S_W - S_B],
\end{aligned}$$

where

$$S_W \equiv \frac{1}{pn(n-1)\sqrt{2}h}\sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n}\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right),$$

and

$$S_B \equiv \frac{1}{p(p-1)n^2\sqrt{2}h}\sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n}\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right). \tag{2.2}$$

The W and B in S_W and S_B stand for within and between, respectively. Intuitively, this is sensible as it compares an $intra-samples$ parameter estimate to an $inter-samples$ parameter estimate. As written in (2.2), S_W and S_B estimate the same parameter, $\int_{-\infty}^{\infty} f^2(x;h)dx$, under H_0, but different parameters under H_A.

Therefore, we may write the test statistic $T_p^{(S)}$, where the upper index S indicates "Simple", as follows:

$$T_p^{(S)} = S_W - S_B, \tag{2.3}$$

where S_W and S_B are defined as in (2.2).

Note that under H_0,

$$E\left(T_p^{(S)}\right) = 0,$$

which illustrates why the same bandwidth is used for $\hat{f}_h(\cdot|i)$ and $\hat{f}_h^{(i)}(\cdot)$. If the same bandwidth were not used, then $E\left(T_p^{(S)}\right)$ would not necessarily be 0 under H_0. Another advantage of using $T_p^{(S)}$ is that we can use the theory of U-statistics to find the asymptotic variance of $T_p^{(S)}$, which will be discussed in Section 2.5, and hence obtain the asymptotic distribution of $T_p^{(S)}$.

2.5 Asymptotic Distribution of $T_p^{(S)}$

In this section, we show that $T_p^{(S)}$ has an asymptotic normal distribution. Since $E\left(T_p^{(S)}\right) = 0$ under H_0, we only need to get an estimate of the variance of $T_p^{(S)}$.

Letting $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, define

$$h_1(\mathbf{x}) = \frac{1}{n(n-1)\sqrt{2}h} \sum_{j=1}^{n} \sum_{l=1, j\neq l}^{n} \phi\left(\frac{x_j - x_l}{\sqrt{2}h}\right),$$

and

$$h_2(\mathbf{x},\mathbf{y}) = \frac{1}{n^2\sqrt{2}h} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi\left(\frac{x_j - y_l}{\sqrt{2}h}\right).$$

With $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})$ for $i = 1, \ldots, p$, we may rewrite

$$S_W = \frac{1}{p} \sum_{i=1}^{p} h_1(\mathbf{X}_i),$$

which is just a mean of i.i.d. random variables, where "i.i.d." stands for independent and identically distributed. We may also rewrite

$$S_B = \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{k\neq i} h_2(\mathbf{X}_i, \mathbf{X}_k),$$

which is a U-statistic. In order to estimate the variance of $T_p^{(S)}$, we find the projection of S_B.

From Serfling (2002) Section 5.3.1, the projection of S_B is defined as

$$\hat{S}_B = \sum_{r=1}^{p} E(S_B|\mathbf{X}_r) - (p-1)E(S_B), \tag{2.4}$$

where $E(S_B) = E[h_2(\mathbf{X}_i, \mathbf{X}_k)] = \theta$.

Define, for each x,

$$h_3(x) = E(h_2(\mathbf{X}_i, \mathbf{X}_k)|\mathbf{X}_k = x) = Eh_2(\mathbf{X}_i, x).$$

We then can write

$$\begin{aligned} E(S_B|\mathbf{X}_r) &= E\left(\frac{1}{p(p-1)}\sum_{i=1}^{p}\sum_{k\neq i} h_2(\mathbf{X}_i, \mathbf{X}_k)|\mathbf{X}_r\right) \\ &= \frac{2}{p}h_3(\mathbf{X}_r) + \frac{p-2}{p}\theta. \end{aligned}$$

Continuing from the projection formula (2.4), we have

$$\begin{aligned} \hat{S}_B &= \sum_{r=1}^{p} E(S_B|\mathbf{X}_r) - (p-1)\theta \\ &= (p-2)\theta + \frac{2}{p}\sum_{i=1}^{p} h_3(\mathbf{X}_i) - (p-1)\theta \\ &= \frac{2}{p}\sum_{i=1}^{p} h_3(\mathbf{X}_i) - \theta, \end{aligned} \tag{2.5}$$

which is a sum of i.i.d. random variables.

Defining $\hat{S} = S_W - \hat{S}_B$, we have

$$\begin{aligned} \text{Var}(\hat{S}) &= \text{Var}\left(S_W - \hat{S}_B\right) \\ &= \text{Var}\left(\frac{1}{p}\sum_{i=1}^{p} h_1(\mathbf{X}_i) - \frac{2}{p}\sum_{i=1}^{p} h_3(\mathbf{X}_i) + \theta\right) \\ &= \frac{1}{p}\text{Var}\left(h_1(\mathbf{X}_i) - 2h_3(\mathbf{X}_i)\right) \\ &\equiv \frac{1}{p}\sigma^2. \end{aligned} \tag{2.6}$$

An important fact is that (see Serfling (2002))

$$\sqrt{p}(\hat{S}_B - S_B) = o_p(1),$$

which implies that $(S_W - S_B)/(\sigma/\sqrt{p})$ has the same asymptotic distribution as $\hat{S}/(\sigma/\sqrt{p})$. Letting "$\xrightarrow{\mathcal{D}}$" stand for convergence in distribution, by the central limit theorem, we know that

$$\frac{\hat{S}}{\sigma/\sqrt{p}} \xrightarrow{\mathcal{D}} N(0,1),$$

and hence

$$\frac{T_p^{(S)}}{\sigma/\sqrt{p}} \xrightarrow{\mathcal{D}} N(0,1).$$

Now we need to find a consistent estimator of $\sigma^2 = p\text{Var}(\hat{S})$.

Apparently, we may use the sample variance of $h_1(\mathbf{X}_i) - 2h_3(\mathbf{X}_i)$, $i = 1, \ldots, p$, to estimate σ^2. Further, from Schucany and Bankson (1989) and Sen (1960), we can obtain an estimate for $h_3(\mathbf{X}_i)$:

$$\begin{aligned}
\hat{h}_3(\mathbf{X}_i) &= \frac{1}{p-1} \sum_{k=1,k\neq i}^{p} h_2(\mathbf{X}_i, \mathbf{X}_k) \\
&= \frac{1}{p-1} \frac{1}{n^2\sqrt{2}h} \sum_{k=1,k\neq i}^{p} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi\left(\frac{X_{ij} - X_{kl}}{\sqrt{2}h}\right)
\end{aligned}$$

for $i = 1, \ldots, p$.

Recall that

$$h_1(\mathbf{X}_i) = \frac{1}{n(n-1)\sqrt{2}h} \sum_{j=1}^{n} \sum_{l=1,j\neq l}^{n} \phi\left(\frac{X_{ij} - X_{il}}{\sqrt{2}h}\right).$$

We then plug in $\hat{h}_3(\mathbf{X}_i)$, and estimate σ^2 by the sample variance of $h_1(\mathbf{X}_i) - 2\hat{h}_3(\mathbf{X}_i)$, $i = 1, \ldots, p$. We call this estimator σ_S^2.

Hence, letting "$\xrightarrow{\mathcal{P}}$" stand for convergence in probability,

$$\sigma_S \xrightarrow{\mathcal{P}} \sigma,$$

as $p \to \infty$.

Define the new standardized test statistic by

$$T^{(S)} = \frac{T_p^{(S)}}{\sigma_S/\sqrt{p}}. \tag{2.7}$$

Under H_0,

$$T^{(S)} \xrightarrow{\mathcal{D}} N(0,1),$$

as $p \to \infty$, by Slutsky's theorem.

One may use the critical values from the standard normal distribution to conduct the test. One will reject H_0 if $T^{(S)} > Z_\alpha$, where Z_α is the upper α percentile of the standard normal distribution. Some commonly used Z_α's are: 1.282, 1.645, and 2.326 corresponding to α equalling 0.10, 0.05, and 0.01, respectively. We use the one-sided test due to the consistency result established in Section 2.7.

2.6 Invariance Property of $T^{(S)}$

The standardized test statistic, $T^{(S)}$, as proposed in (2.7), is invariant to both location and scale under the null. Due to this property, we don't have to pick specific location and scale parameters for a distribution in the simulation studies. For example, when we want to study the behavior under the null when all the distributions are from the student t, it doesn't matter if we pick t_3 or $t_3/\sqrt{3}$, which has standard deviation equal to 1, because of the scale invariance of the test statistic.

To prove the invariance property, note first that if the original data are $\mathbf{X}$, we may write a new data set as $\mathbf{Y} = a\mathbf{X} + b$, where a is a positive constant indicating the scale difference, and b is the location difference. In this case, the sample standard deviations change from s_i to $a \cdot s_i$ for each $i = 1, \ldots, p$, and the pooled sample standard deviation changes to $a \cdot s_{pool}$. Therefore, defining h_Y to be the bandwidth for data set $\mathbf{Y}$, we have $h_Y = ah$.

Recall from Section 2.4, we may write

$$T_p^{(S)}(\mathbf{X}) \;=\; S_W(\mathbf{X}) - S_B(\mathbf{X}),$$

where

$$\begin{aligned} S_W(\mathbf{X}) &= \frac{1}{pn(n-1)\sqrt{2}h} \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{l=1, l\neq j}^{n} \phi\left(\frac{X_{ij} - X_{il}}{\sqrt{2}h}\right) \\ S_B(\mathbf{X}) &= \frac{1}{p(p-1)n^2\sqrt{2}h} \sum_{i=1}^{p} \sum_{k=1, k\neq i}^{p} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi\left(\frac{X_{ij} - X_{kl}}{\sqrt{2}h}\right), \end{aligned}$$

for the original data set $\mathbf{X}$.

Then, for $\mathbf{Y}$, we have

$$\begin{aligned} S_W(\mathbf{Y}) &= \frac{1}{pn(n-1)\sqrt{2}h_Y} \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{l=1, l\neq j}^{n} \phi\left(\frac{Y_{ij} - Y_{il}}{\sqrt{2}h_Y}\right) \\ &= \frac{1}{pn(n-1)\sqrt{2}ah} \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{l=1, l\neq j}^{n} \phi\left(\frac{(aX_{ij} + b) - (aX_{il} + b)}{\sqrt{2}ah}\right) \\ &= \frac{1}{a} \frac{1}{pn(n-1)\sqrt{2}h} \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{l=1, l\neq j}^{n} \phi\left(\frac{X_{ij} - X_{il}}{\sqrt{2}h}\right) \\ &= \frac{1}{a} S_W(\mathbf{X}). \end{aligned}$$

Similarly,

$$S_B(\mathbf{Y}) = \frac{1}{a} S_B(\mathbf{X}).$$

Therefore,

$$T_p^{(S)}(\mathbf{Y}) = \frac{1}{a} T_p^{(S)}(\mathbf{X}). \tag{2.8}$$

For the variance estimate, define the standard deviation for the new data set $\mathbf{Y}$ as

$\sigma_S(\mathbf{Y})$:

$$\begin{aligned}
\sigma_S(\mathbf{Y}) &= \sqrt{\widehat{\text{Var}}\left(h_1(\mathbf{Y}_i) - 2\hat{h}_3(\mathbf{Y}_i)\right)} \\
&= \sqrt{\widehat{\text{Var}}\left(h_1(a\mathbf{X}_i + b) - 2\hat{h}_3(a\mathbf{X}_i + b)\right)} \\
&= \sqrt{\widehat{\text{Var}}\left(\frac{1}{a}h_1(\mathbf{X}_i) - \frac{2}{a}\hat{h}_3(\mathbf{X}_i)\right)} \\
&= \frac{1}{a}\sqrt{\widehat{\text{Var}}\left(h_1(\mathbf{X}_i) - 2\hat{h}_3(\mathbf{X}_i)\right)} \\
&= \frac{1}{a}\sigma_S(\mathbf{X}). \qquad (2.9)
\end{aligned}$$

Combining the results from (2.8) and (2.9), it follows that

$$\begin{aligned}
T^{(S)}(\mathbf{Y}) &= \frac{T_p^{(S)}(\mathbf{Y})}{\sigma_S(\mathbf{Y})/\sqrt{p}} \\
&= \frac{\frac{1}{a}T_p^{(S)}(\mathbf{X})}{\frac{1}{a}\sigma_S(\mathbf{X})/\sqrt{p}} \\
&= T^{(S)}(\mathbf{X}).
\end{aligned}$$

2.7 Power Analysis

A test is consistent if the power against a fixed alternative has the limit of 1 as the sample size goes to infinity, while the size of the test is fixed. In this section, we find conditions under which our KDE-based test is consistent. We will show some plots of power analysis in the section of simulation studies.

In Section 2.7.1 and 2.7.2, we assume the alternative hypothesis is true and each of the p data sets comes from one of two different distributions, say f and g. In particular, it is assumed that each data set comes from density g with probability ρ and from f with probability $1 - \rho$, where ρ is a constant between 0 and 1. Once a density (f or g) is selected, a random sample of size n is taken from the selected density. It follows that the unconditional distribution of X_{ij} is as follows:

$$X_{ij} \sim m(x) = \rho g(x) + (1 - \rho)f(x), \qquad (2.10)$$

for $i = 1, \ldots, p$, $j = 1, \ldots, n$.

It is worthwhile noting that under this model (assuming $f \neq g$), X_{ij} and X_{kl} are independent iff $i \neq k$, where $i,\ k = 1, \ldots, p$, $j,\ l = 1, \ldots, n$.

In Section 2.7.3, we show that under a general alternative, when the distribution of X_{ij} is a mixture of a countable number of distributions, our test is still consistent.

2.7.1 Consistency for the "Easy" Case

By the "easy" case, we mean that ρ in (2.10) is fixed. In Section 2.7.2, we will consider a case where $\rho \to 0$, as $p \to \infty$. In the chapter on simulation studies, we choose $\rho = 0.1$ or 0.2. We also assume h is fixed as $p \to \infty$.

Let's first define the following notations:

$$\begin{aligned} g(x;h) &= \frac{1}{h}\int_{-\infty}^{\infty} \phi\left(\frac{x-u}{h}\right) g(u)du, \\ f(x;h) &= \frac{1}{h}\int_{-\infty}^{\infty} \phi\left(\frac{x-u}{h}\right) f(u)du. \end{aligned}$$

Proposition 2.1:

Suppose $\int_{-\infty}^{\infty} \left(f(x) - g(x)\right)^2 dx > 0$. Define

$$\mu_A \equiv E\left(T_p^{(S)} | H_A\right) \tag{2.11}$$

to be the expected value under H_A. Then,

$$\begin{aligned} \mu_A &= \rho(1-\rho)\sqrt{2}h \int_{-\infty}^{\infty} \left(f(x;h) - g(x;h)\right)^2 dx \qquad (2.12) \\ &> 0. \end{aligned}$$

Proof of Proposition 2.1:

Recall that our test statistic is: $T_p^{(S)} = S_W - S_B$, with S_W and S_B defined as in (2.2).

Let E_i be the event that data set i comes from density g. Then

$$\begin{aligned}
E(S_W|H_A) &= \frac{1}{\sqrt{2}h}E\left(\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right)\middle|H_A\right) \\
&= \frac{\rho}{\sqrt{2}h}E\left(\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right)\middle|E_i\right)+\frac{1-\rho}{\sqrt{2}h}E\left(\phi\left(\frac{X_{ij}-X_{il}}{\sqrt{2}h}\right)\middle|E_i^c\right) \\
&= \frac{\rho}{\sqrt{2}h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(\frac{x-y}{\sqrt{2}h}\right)g(x)g(y)dxdy \\
&\quad +\frac{1-\rho}{\sqrt{2}h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(\frac{x-y}{\sqrt{2}h}\right)f(x)f(y)dxdy.
\end{aligned}$$

In the summation defining S_B, $i \neq k$ and hence

$$\begin{aligned}
E(S_B|H_A) &= \frac{1}{\sqrt{2}h}E\left(\phi\left(\frac{X_{ij}-X_{kl}}{\sqrt{2}h}\right)\middle|H_A\right) \\
&= \frac{1}{\sqrt{2}h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(\frac{x-y}{\sqrt{2}h}\right)m(x)m(y)dxdy \\
&= \frac{1}{\sqrt{2}h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(\frac{x-y}{\sqrt{2}h}\right)(\rho g(x)+(1-\rho)f(x))\,(\rho g(y)+(1-\rho)f(y))\,dxdy.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mu_A &= E(S_W|H_A)-E(S_B|H_A) \\
&= \frac{\rho(1-\rho)}{\sqrt{2}h}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi\left(\frac{x-y}{\sqrt{2}h}\right)(f(x)-g(x))(f(y)-g(y))dxdy \\
&\quad \text{(Convolution of two normal functions)} \\
&= \rho(1-\rho)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{1}{h}\phi\left(\frac{z-x}{h}\right)\frac{1}{h}\phi\left(\frac{z-y}{h}\right)dz\right) \\
&\quad \cdot(f(x)-g(x))(f(y)-g(y))dxdy \\
&= \rho(1-\rho)\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{1}{h}\phi\left(\frac{z-x}{h}\right)(f(x)-g(x))dx\right)^2dz \\
&= \rho(1-\rho)\int_{-\infty}^{\infty}\left(f(z;h)-g(z;h)\right)^2dz \\
&> 0, \qquad (2.13)
\end{aligned}$$

by the condition that $\int_{-\infty}^{\infty}\left(f(x)-g(x)\right)^2dx > 0$.

Note that our test statistic in practice has the following form:

$$T^{(S)} = \frac{T_p^{(S)}}{\sigma_S/\sqrt{p}}.$$

It has been proved that the asymptotic distribution of $T^{(S)}$ is the standard normal when H_0 is true. For a test with significance level 0.05, normally, an absolute value of the test statistic greater than 1.96 will result in rejection of the null hypothesis for a two-sided test. However, due to *Proposition 2.1*, a one-sided test is appropriate since only positive values of $T^{(S)}$ favor the alternative.

In the simulation studies, we also show some plots illustrating the fact that the test statistic's distribution shifts to the right under H_A. Therefore, in practice, we will reject the null hypothesis whenever the test statistic, $T^{(S)}$, is greater than Z_α, say, 1.645 at significance level 0.05. By using a one-sided test, we enhance the power of the test.

Recall that σ_S^2 in Section 2.5 estimates the asymptotic variance of $T_p^{(S)}$ under H_0. Define under H_A,

$$\sigma_A^2 \equiv \text{Var}(T_p^{(S)}|H_A) = \text{Var}(h_1(\boldsymbol{X}_i) - 2h_3(\boldsymbol{X}_i)|H_A).$$

Then, σ_S is a consistent estimator of σ_A under H_A.

Define $\delta(\boldsymbol{X}_i) = h_1(\boldsymbol{X}_i) - 2h_3(\boldsymbol{X}_i)$. Under H_A, $\sigma_A^2 = \text{Var}(\delta(\boldsymbol{X}_i))$. Use the iterated expectation formula for variance, i.e.,

$$\text{Var}(\delta(\boldsymbol{X}_i)) = \text{Var}[E(\delta(\boldsymbol{X}_i)|\mathcal{F})] + E[\text{Var}(\delta(\boldsymbol{X}_i)|\mathcal{F})],$$

where $\mathcal{F}$ denotes "random distribution," f or g with respective probabilities $1-\rho$ and ρ.

Using calculations similar to those in the proof of *Proposition 2.1*, we can show that

$$\text{Var}[E(\delta(\boldsymbol{X}_i)|\mathcal{F})] = \rho(1-\rho)\left[\int f^2(u;h)\,du - \int g^2(u;h)\,du\right]^2, \tag{2.14}$$

and obviously

$$E[\text{Var}(\delta(\boldsymbol{X}_i)|\mathcal{F})] = \rho\text{Var}(\delta(\boldsymbol{X}_i)|g) + (1-\rho)\text{Var}(\delta(\boldsymbol{X}_i)|f). \tag{2.15}$$

Now, Var$(\delta(\boldsymbol{X}_i)|f)$ is the asymptotic variance of $T_p^{(S)}$ under the null hypothesis. It is the quantity estimated by σ_S when H_0 is true.

Expressions (2.14) and (2.15) allow us to see how the variance of the test statistic changes under various kinds of alternatives. For example, suppose f is $N(0,1)$ and g is $N(\mu,1)$. Then Var$(\delta(\boldsymbol{X}_i)|g) =$ Var$(\delta(\boldsymbol{X}_i)|f)$ and $\int f^2(u;h)\,du = \int g^2(u;h)\,du$, which means that Var$(\delta(\boldsymbol{X}_i))$ is the same as it is under the null hypothesis.

Next, we show that our test is consistent.

Theorem 2.2:

For a one-sided test of level α, define Z_α to be the $(1-\alpha)100\%$ quantile of the standard normal distribution. Then

$$\Pr(T^{(S)} > Z_\alpha | H_A) \to 1, \tag{2.16}$$

as $p \to \infty$.

Proof of Theorem 2.2:

Define under H_A,

$$\sigma_A^2 \equiv \lim_{p\to\infty} p\,\text{Var}\left(T_p^{(S)}\right) > 0.$$

Recall that we have defined $\mu_A = E\left(T_p^{(S)}|H_A\right)$ in (2.11). Therefore, by the Central Limit Theorem, we have

$$Z = \frac{\sqrt{p}\left(T_p^{(S)} - \mu_A\right)}{\sigma_A} \xrightarrow{\mathcal{D}} N(0,1),$$

as $p \to \infty$.

The power of the test equals:

$$\begin{aligned}
\Pr(T^{(S)} > Z_\alpha | H_A) &= \Pr\left(\frac{\sqrt{p}\, T_p^{(S)}}{\sigma_S} > Z_\alpha \middle| H_A \right) \\
&= \Pr\left(\frac{\sqrt{p}\left(T_p^{(S)} - \mu_A\right)}{\sigma_S} + \frac{\sqrt{p}\,\mu_A}{\sigma_S} > Z_\alpha \middle| H_A \right) \\
&= \Pr\left(\frac{\sqrt{p}\left(T_p^{(S)} - \mu_A\right)}{\sigma_A} \cdot \frac{\sigma_A}{\sigma_S} + \frac{\sqrt{p}\,\mu_A}{\sigma_S} > Z_\alpha \middle| H_A \right) \\
&= \Pr\left(Z > Z_\alpha \cdot \frac{\sigma_S}{\sigma_A} - \frac{\sqrt{p}\,\mu_A}{\sigma_A} \middle| H_A \right). \qquad (2.17)
\end{aligned}$$

Recall from *Proposition 2.1* , $\mu_A > 0$ under H_A. Hence,

$$\frac{\sqrt{p}\,\mu_A}{\sigma_A} \to \infty,$$

when $p \to \infty$.

Also recall that σ_S converges in probability to σ_A as $p \to \infty$. Therefore,we have

$$Z_\alpha \cdot \frac{\sigma_S}{\sigma_A} - \frac{\sqrt{p}\,\mu_A}{\sigma_A} \to -\infty,$$

and hence the proof is complete.

2.7.2 Consistency for the "Harder" Case

The "harder" case uses the same model, but now we suppose that ρ converges to zero as p goes to infinity . Intuitively, when ρ is a very small number close to zero, it will be very hard to detect that there are two different distributions. It would be ideal for us to find out how small this ρ could be and still obtain a consistent test.

Recall that in the "easy" case (2.17), in order to get a consistent test, we require

$$\frac{\sqrt{p}\,\mu_A}{\sigma_A} \to \infty,$$

as $p \to \infty$.

We consider the same limit of $\sqrt{p}\,\mu_A$ in the "harder" case. From (2.12), we have

$$\sqrt{p}\,\mu_A \;=\; \sqrt{p}\,\rho(1-\rho)\int_{-\infty}^{\infty}\left(f(x;h)-g(x;h)\right)^2 dx.$$

We need $\sqrt{p}\,\mu_A \to \infty$, as $p \to \infty$, and hence $\sqrt{p}\rho$ should be unbounded. It suffices to take $\rho \sim Cp^{-\frac{1}{2}+\epsilon}$, for some positive ϵ and positive constant C. We also need $-\frac{1}{2}+\epsilon < 0$, i.e. $0 < \epsilon < \frac{1}{2}$, because $\rho \to 0$ as $p \to \infty$.

In other words, in order to get consistency in the "harder" case, we need the following condition:

$$\rho \sim Cp^{-\frac{1}{2}+\epsilon},$$

where $0 < \epsilon < \frac{1}{2}$, and C is some positive constant.

2.7.3 Consistency for a General Alternative

Assume that the alternative is true and the density for any given small data set is drawn from a countable collection $\{f_1, f_2, \ldots\}$ in such a way that $P(f_r) = \rho_r$, $r = 1, 2, \ldots$. Once a density is selected, then a random sample of size n is drawn from that density. For the i-th data set, we have

$$\begin{aligned} E\phi\left(\frac{X_{ij}-X_{i\ell}}{\sqrt{2}h}\right) &= \sum_{r=1}^{\infty}\rho_r E\left[\phi\left(\frac{X_{ij}-X_{i\ell}}{\sqrt{2}h}\right)\middle| f_r\right] \\ &= \sum_{r=1}^{\infty}\rho_r \int\int \phi\left(\frac{x-y}{\sqrt{2}h}\right) f_r(x) f_r(y)\,dxdy. \end{aligned}$$

If $i \neq k$, then

$$E\phi\left(\frac{X_{ij}-X_{k\ell}}{\sqrt{2}h}\right) = \int\int \phi\left(\frac{x-y}{\sqrt{2}h}\right) m(x)m(y)\,dxdy,$$

where

$$m(x) = \sum_{r=1}^{\infty}\rho_r f_r(x).$$

Analogous to (2.13),

$$\mu_A = \frac{1}{\sqrt{2}h}\int\int \phi\left(\frac{x-y}{\sqrt{2}h}\right)\left[\sum_{r=1}^{\infty}\rho_r f_r(x) f_r(y)\sum_{s=1}^{\infty}\rho_s - \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\rho_r\rho_s f_r(x) f_s(y)\right] dxdy$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2}h}\int\int \phi\left(\frac{x-y}{\sqrt{2}h}\right)\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\rho_r\rho_s f_r(x)\left[f_r(y)-f_s(y)\right]\,dxdy \\
&= \frac{1}{\sqrt{2}h}\int\int \phi\left(\frac{x-y}{\sqrt{2}h}\right)\sum_{r=1}^{\infty}\sum_{s>r}\rho_r\rho_s (f_r(x)-f_s(x))(f_r(y)-f_s(y))\,dxdy \\
&= \sum_{r=1}^{\infty}\sum_{s>r}\rho_r\rho_s\int (f_r(x;h)-f_s(x;h))^2\,dx \\
&> 0,
\end{aligned}
$$

where, for all r,

$$f_r(x;h) = \frac{1}{\sqrt{2}h}\int \phi\left(\frac{x-u}{\sqrt{2}h}\right) f_r(u)\,du.$$

Therefore,

$$\frac{\sqrt{p}\,\mu_A}{\sigma_A} \to \infty,$$

when $p \to \infty$, and hence we complete the proof of consistency for a general alternative.

CHAPTER III

SIMULATION

In this chapter, we investigate the power of the test by using simulations. We perform simulations under several different situations, with different p's, different n's, different ρ's, and different underlying distributions. We set three commonly used levels of significance, α, to be 0.01, 0.05 and 0.10. In Sections 3.1 and 3.2, we list the possible settings of H_0 and H_A for the simulation studies. In Section 3.3, we provide critical values for the test and compare the true levels of the test under the null with the nominal significance levels. This is followed by the results in Section 3.4, which include tables and plots from the simulation. In Section 3.5, we compare our test with some existing methods. R functions, named **TS**, for computing $T^{(S)}$ are available in Appendix III.

3.1 Settings for H_0

There are four different settings for H_0:

1. Normal case: Standard normal distributions, $N(0, 1)$.
2. t_3 case: t_3 distributions with mean 0 and standard deviation $\sqrt{3}$.
3. Mixed case: $\frac{1}{2}\Phi(x-2) + \frac{1}{2}\Phi(x+2)$, where $\Phi(\cdot)$ is the CDF of a standard normal distribution.
4. Gamma case: $\mathrm{gamma}(\mathrm{shape} = 3, \mathrm{scale} = 1)$.

These cases represent four different scenarios. Normal case represents symmetric distributions with short tails; the t_3 case represents symmetric distributions with heavy tails; the mixed case represents a symmetric bimodal mixture of normal distributions; and

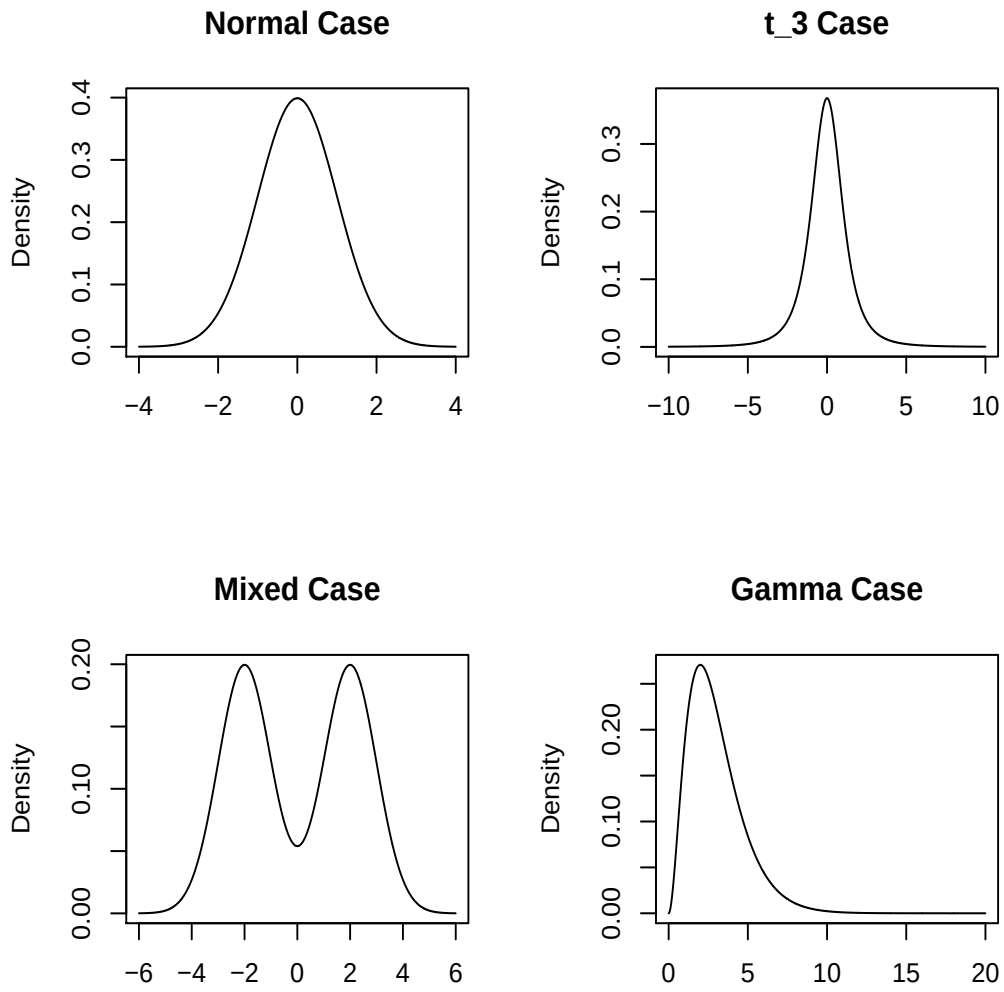

Figure 3: The four distributions for the four different settings under H_0.

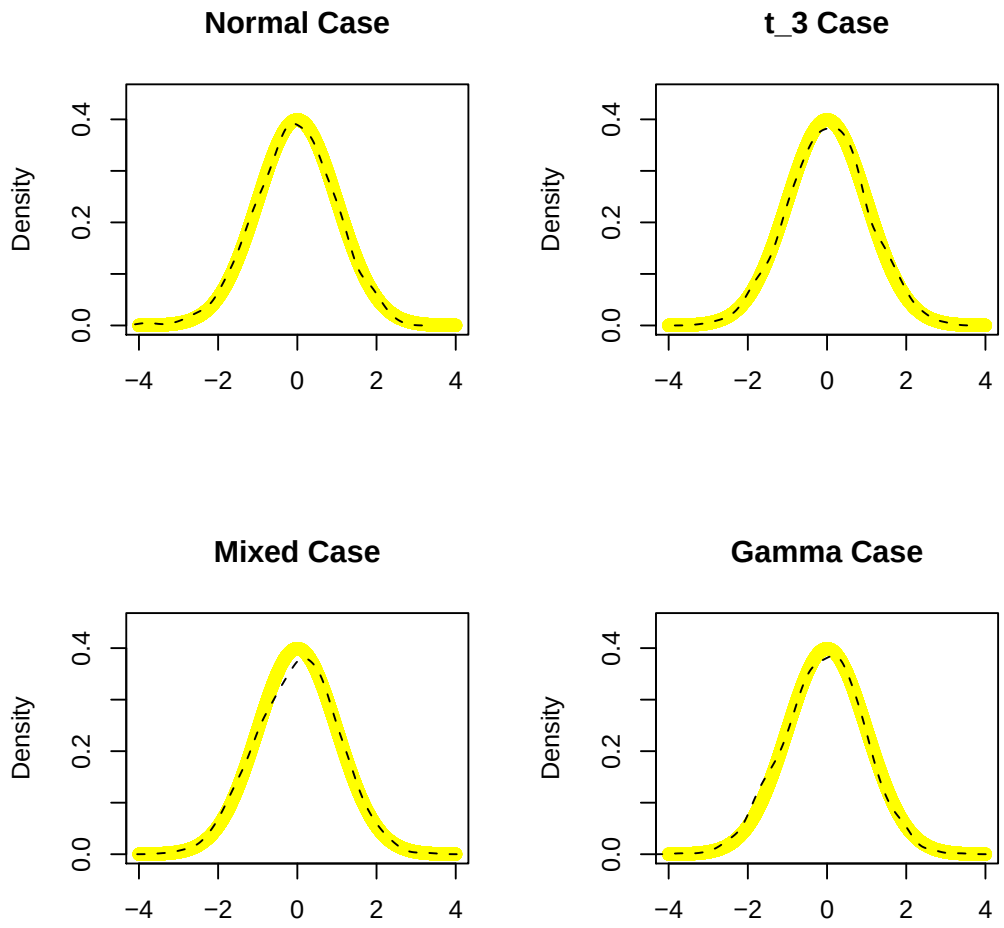

Figure 4: Plots of kernel density estimates of the density of $T^{(S)}$ under four different settings of H_0, with p=1000, n=5. The yellow lines are the standard normal distribution curves. Each dashed black line is the kernel density estimate from 1000 simulated data sets.

the gamma case represents skewed distributions. These four distributions are shown in Figure 3.

Figure 4 shows estimates of the density of the standardized test statistic $T^{(S)}$ under the null hypothesis for the four cases, when $p = 1000$, $n = 5$. From the plots, we can see that the kernel density estimate of the density of $T^{(S)}$ is quite close to the standard normal distribution for each case.

3.2 Settings for H_A

Define $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})$, $i = 1, \ldots, p$.

For the Normal case and t_3 case, we set the alternative to be one of the following:
When $\rho = 0.1$ or 0.2,
(1) $\rho \cdot 100\%$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ have a "location difference" from the rest of the data sets;
(2) $\rho \cdot 100\%$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ have a "scale difference" from the rest of the data sets;
(3) $\rho \cdot 100\%$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ have a "shape difference" from the rest of the data sets. By "shape difference", we mean that the alternative distribution is not just a shifted and/or rescaled version of the null distribution. Note that when testing location difference, we keep the scale the same, and vice versa.

For the Mixed case, we only set the alternative to be a "shape difference". For the Gamma case, the alternative is set to be another gamma distribution with different scale parameters.

Suppose that $\rho \cdot 100\%$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ are from distribution g under the alternative, then detailed settings are as follows:

1. Normal case:

   - $H_A(1)$: Location difference: $g = N(1, 1)$.
   - $H_A(2)$: Scale difference: $g = N(0, 2^2)$.

- $H_A(3)$: Shape difference: g = exp(1)-1.

2. t_3 case:

   - $H_A(4)$: Location difference: $g = t_3 + 1$.
   - $H_A(5)$: Scale difference: $g = t_{30}$, with standard deviation $\sqrt{\frac{15}{14}}$.
   - $H_A(6)$: Shape difference: g = exp(1)-1.

3. Mixed case:

   - $H_A(7)$: Shape difference: g=exp(1)-1.

4. Gamma case:

   - $H_A(8)$: Location and scale differences: $g = \text{gamma}(\text{shape} = 3, \text{scale} = \beta_i)$, where $\beta_i \sim \text{gamma}(\text{shape} = 2, \text{scale} = 0.5), i = 1, \ldots, p; j = 1, \ldots, n$.

Figure 5 and Figure 6 plot these eight densities (g in red) versus the corresponding nulls (f in black). These plots visually illustrate how g is different from f. It is noticed that for some of the alternatives, the curves are not significantly different from each other, which could lead to low powers in the simulation study. For example, $H_A(5)$ has both curves so close to each other, so that it will be very hard to detect the differences between the null and the alternative.

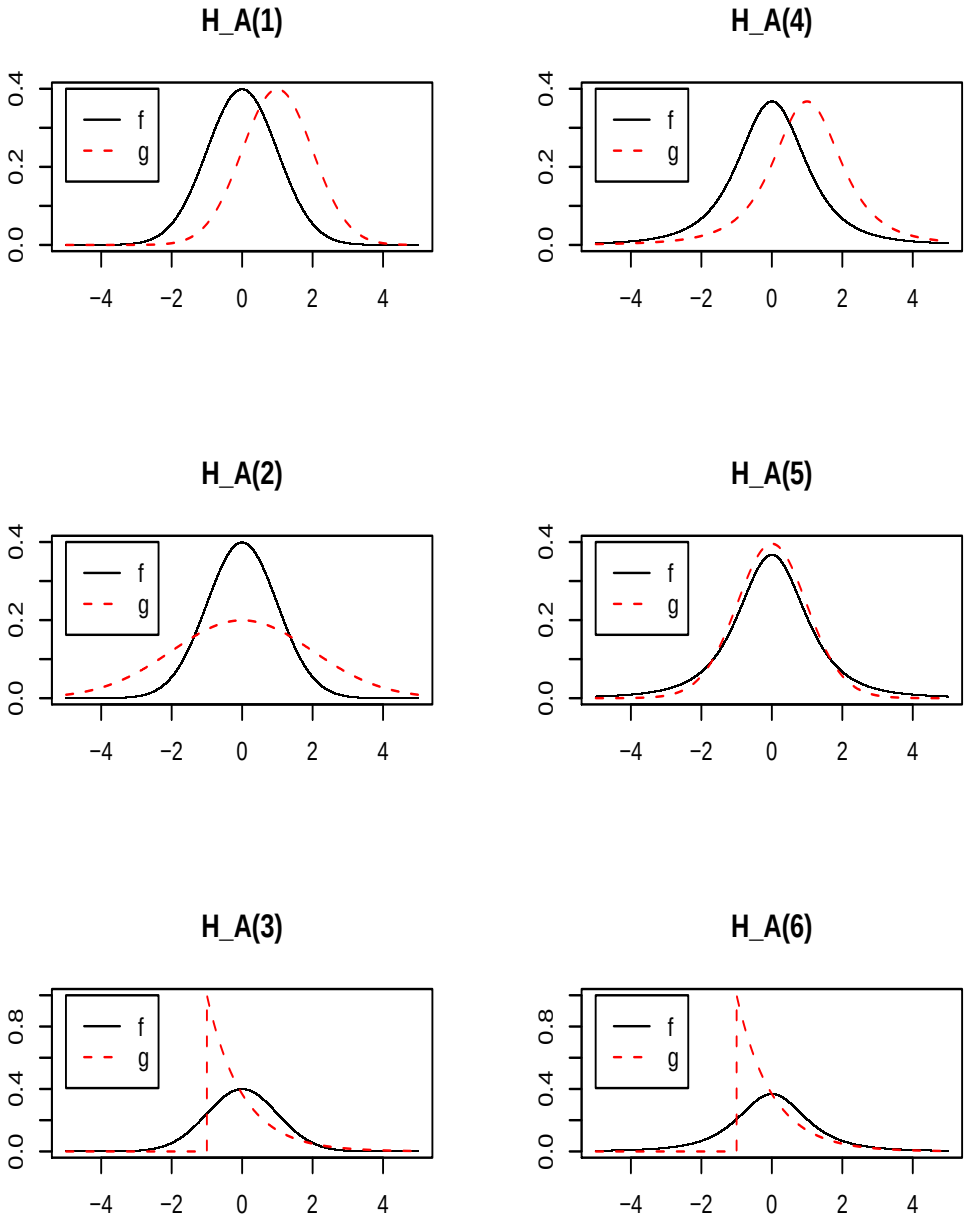

Figure 5: Plots of densities (g) under the alternatives $H_A(1)$, $H_A(2)$, $H_A(3)$, $H_A(4)$, $H_A(5)$, and $H_A(6)$. The black solid curve (f) is the density under the null, which is $N(0, 1^2)$ for $H_A(1)$, $H_A(2)$, $H_A(3)$ and t_3 for $H_A(4)$, $H_A(5)$, $H_A(6)$.

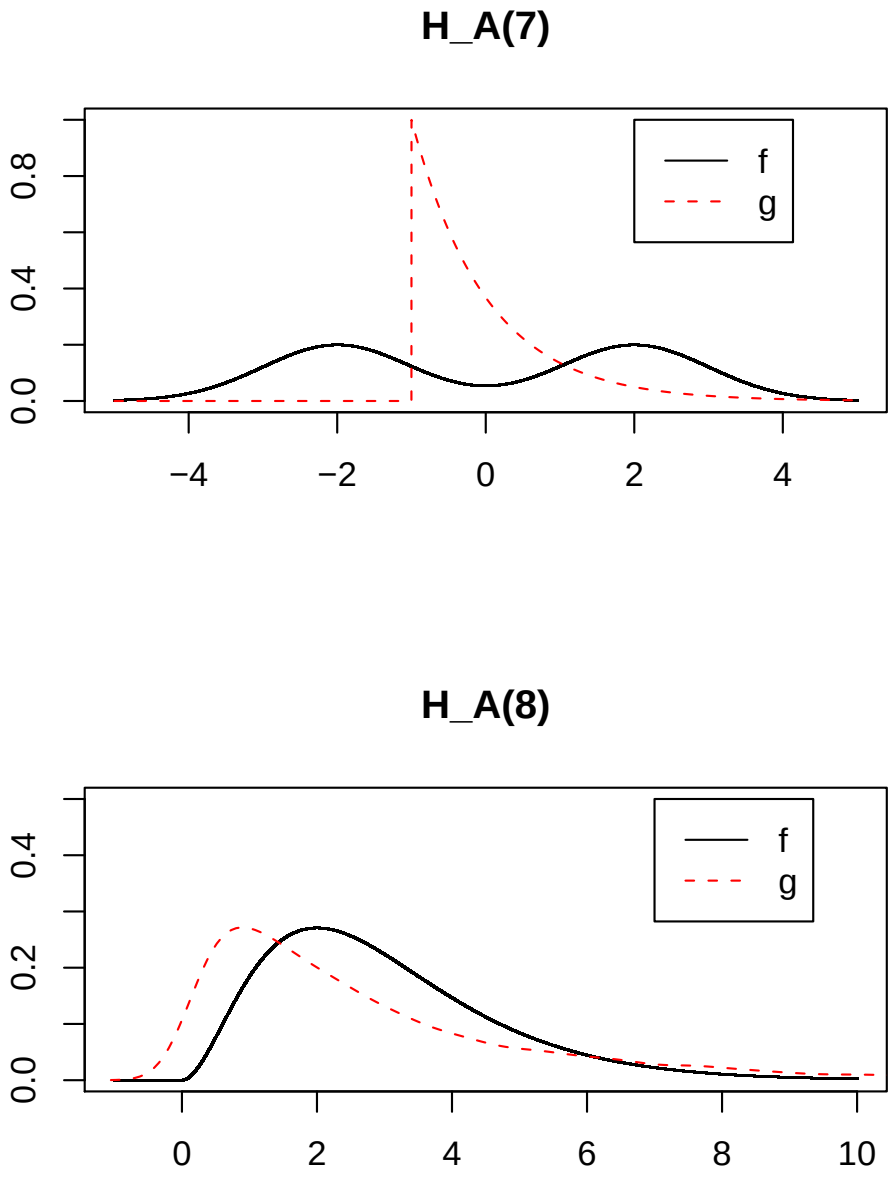

Figure 6: Plots of densities (g) under the alternatives $H_A(7)$ and $H_A(8)$. The black solid curve (f) is the density under the null, whose distribution is $\frac{1}{2}\Phi(x-2)+\frac{1}{2}\Phi(x+2)$ for $H_A(7)$. For $H_A(8)$, the null distribution is $\mathrm{gamma}(\mathrm{shape}=3, \mathrm{scale}=1)$, the red curve ($g$) is a kernel density estimate of the alternative when $p=1000$, $n=5$.

3.3 Critical Value and Type I Error

Due to the asymptotic normality property of the test statistic, we may use $(1-\alpha)\cdot 100\%$ quantiles from the standard normal distribution as critical values. The corresponding values for the three most popular α's are in Table 1. We reject H_0 if the value of $T^{(S)}$ is greater than Z_α.

Table 1: The three most commonly used critical values from the standard normal distribution. The level of the test is α, and Z_α is the critical value.

α	0.01	0.05	0.10
Z_α	2.3263	1.6449	1.2816

To ensure validity of the test, the probability of a type I error should be close to the nominal significance level α. Therefore, we run simulations to check whether the type I error probabilities of our test are close to the nominal significance levels. Some examples of comparison of the true significance level and nominal significance levels are shown in Table 2 to Table 5 for each null case. It shows that our test has an empirical significance level close to the nominal one.

Table 2: Comparison of true and nominal significance levels. The number in each cell is the empirical probability of type I error when Z_α is used as critical value. H_0 : Normal case.

		Prob (Type I Error)		
n	p	α =0.01	α =0.05	α =0.10
2	100	0.009	0.052	0.099
	500	0.010	0.045	0.093
	1000	0.014	0.054	0.103
	5000	0.012	0.057	0.104
3	100	0.009	0.036	0.076
	500	0.012	0.044	0.094
	1000	0.009	0.048	0.105
	5000	0.010	0.047	0.089
5	100	0.005	0.040	0.091
	500	0.007	0.045	0.088
	1000	0.015	0.044	0.094
	5000	0.008	0.049	0.082
10	100	0.005	0.038	0.088
	500	0.004	0.047	0.081
	1000	0.005	0.046	0.085
	5000	0.007	0.046	0.091

Table 3: Comparison of true and nominal significance levels. The number in each cell is the empirical probability of type I error when Z_α is used as critical value. $H_0 : t_3$ case.

		Prob (Type I Error)		
n	p	α =0.01	α =0.05	α =0.10
2	100	0.007	0.033	0.076
	500	0.007	0.049	0.103
	1000	0.013	0.041	0.085
	5000	0.012	0.044	0.090
3	100	0.007	0.039	0.089
	500	0.009	0.042	0.081
	1000	0.007	0.042	0.092
	5000	0.008	0.045	0.103
5	100	0.006	0.039	0.076
	500	0.007	0.041	0.085
	1000	0.012	0.058	0.114
	5000	0.011	0.055	0.110
10	100	0.006	0.038	0.078
	500	0.006	0.040	0.102
	1000	0.013	0.058	0.121
	5000	0.012	0.056	0.108

Table 4: Comparison of true and nominal significance levels. The number in each cell is the empirical probability of type I error when Z_α is used as critical value. H_0 : Mixed case.

		Prob (Type I Error)		
n	p	α =0.01	α =0.05	α =0.10
2	100	0.011	0.041	0.081
	500	0.013	0.055	0.104
	1000	0.012	0.058	0.117
	5000	0.010	0.056	0.106
3	100	0.003	0.029	0.079
	500	0.014	0.053	0.116
	1000	0.012	0.048	0.092
	5000	0.011	0.051	0.102
5	100	0.007	0.034	0.079
	500	0.007	0.045	0.092
	1000	0.008	0.051	0.101
	5000	0.011	0.052	0.103
10	100	0.004	0.037	0.082
	500	0.007	0.044	0.084
	1000	0.008	0.047	0.083
	5000	0.008	0.048	0.093

Table 5: Comparison of true and nominal significance levels. The number in each cell is the empirical probability of type I error when Z_α is used as critical value. H_0 : Gamma case.

		Prob (Type I Error)		
n	p	α =0.01	α =0.05	α =0.10
2	100	0.004	0.044	0.085
	500	0.010	0.041	0.088
	1000	0.007	0.050	0.093
	5000	0.011	0.051	0.103
3	100	0.004	0.036	0.080
	500	0.009	0.042	0.083
	1000	0.009	0.045	0.089
	5000	0.010	0.049	0.101
5	100	0.003	0.036	0.075
	500	0.007	0.052	0.102
	1000	0.007	0.042	0.107
	5000	0.009	0.048	0.099
10	100	0.004	0.034	0.076
	500	0.006	0.044	0.098
	1000	0.007	0.044	0.089
	5000	0.007	0.046	0.092

3.4 Results: Empirical Powers from Simulation

In this section, we provide the empirical powers for each case under different alternatives. The number for each simulation is 1000. The power was calculated as the percentage of the tests among 1000 simulations that were rejected according to the critical values.

3.4.1 Distributions For Different Alternatives

As mentioned in Section 2.7.1, a one-sided test is adequate in our test, which results in a more powerful test. Figures 7 to 14 all show that the distribution of $T^{(S)}$ under H_A is shifted to the right of the standard normal distribution. We choose $p = 1000$, $n = 5$ and $\rho = 0.1$ for each alternative. Although for $H_A(3)$, $H_A(5)$ and $H_A(6)$, the shifts to the right are small, making the power low, they still confirm the fact that a one-sided test is the right thing to do.

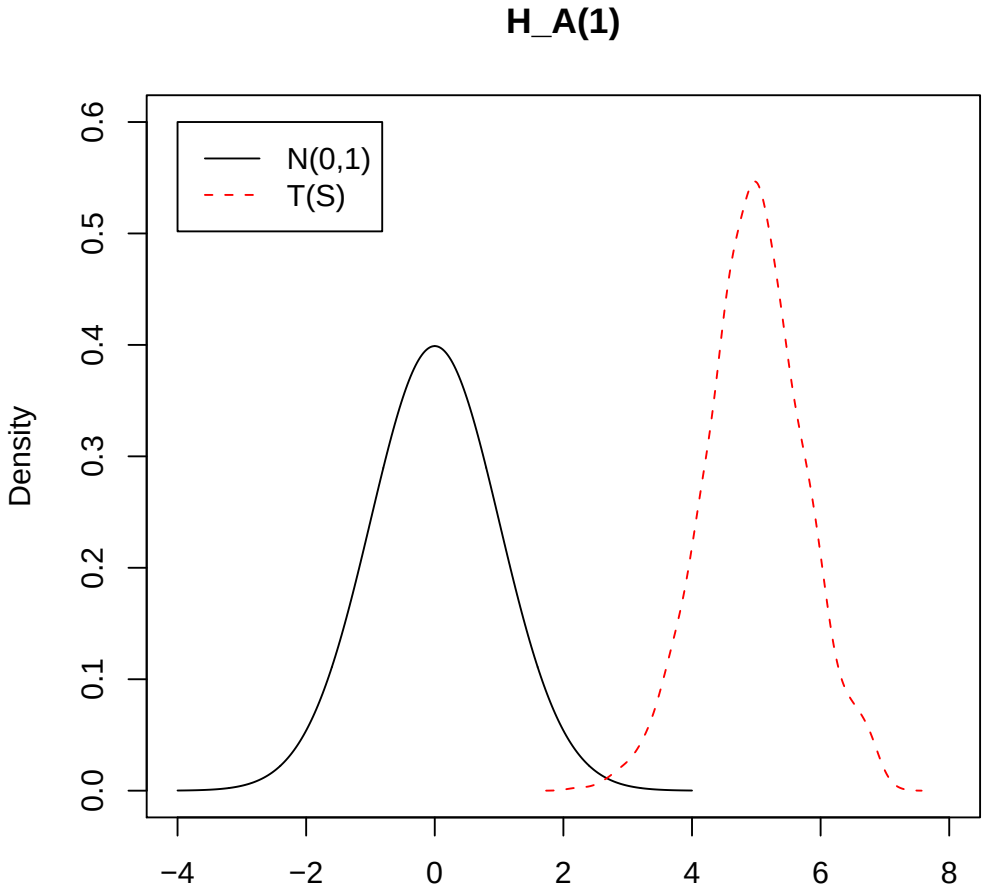

Figure 7: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(1)$, when p=1000, n=5, and $\rho = 0.1$.

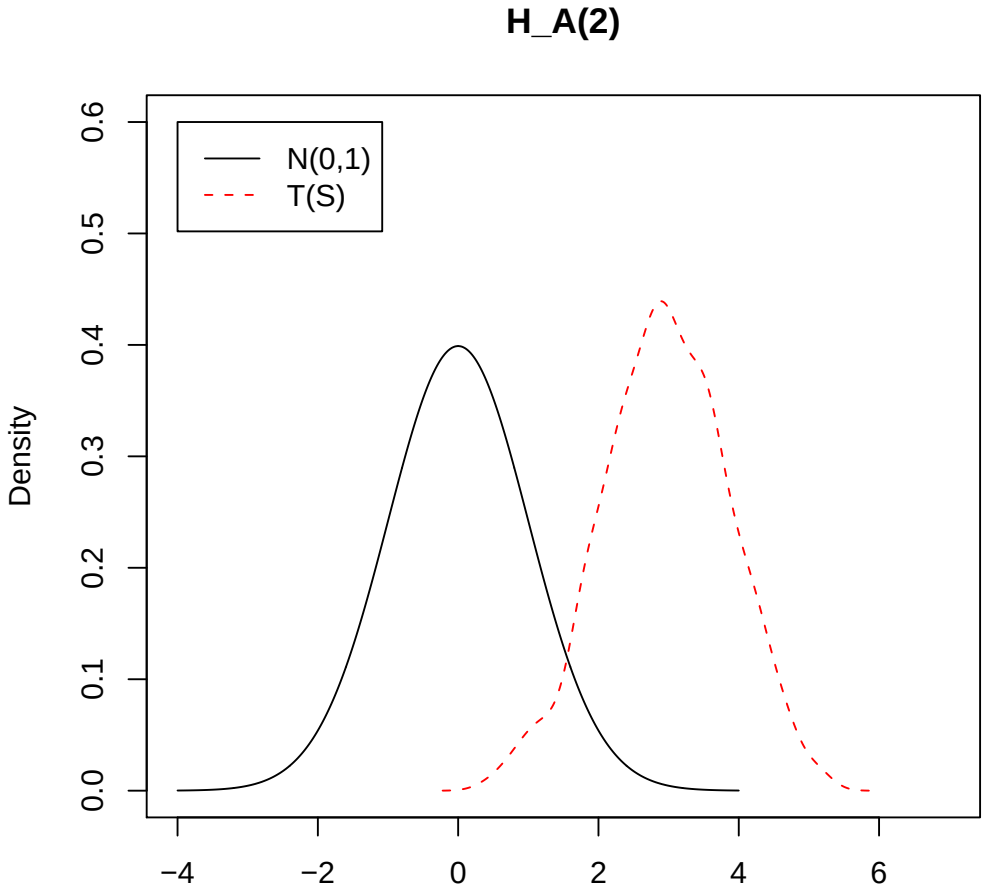

Figure 8: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(2)$, when p=1000, n=5, and $\rho = 0.1$.

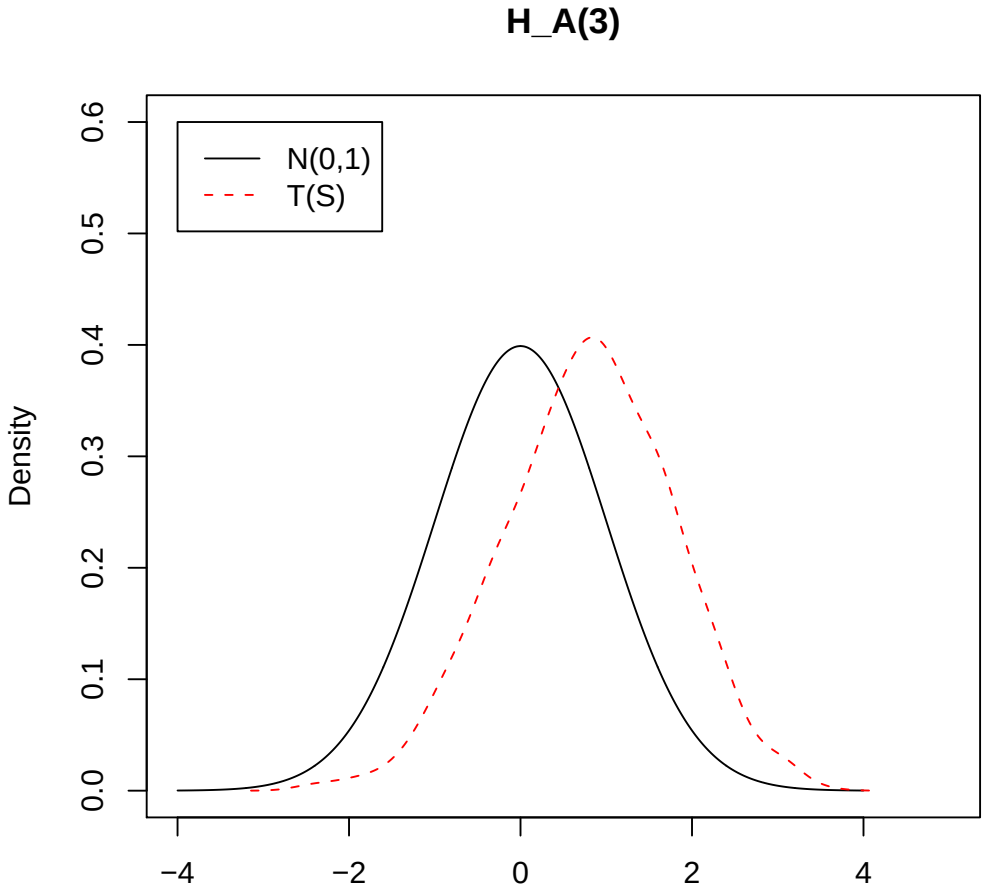

Figure 9: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(3)$, when p=1000, n=5, and $\rho = 0.1$.

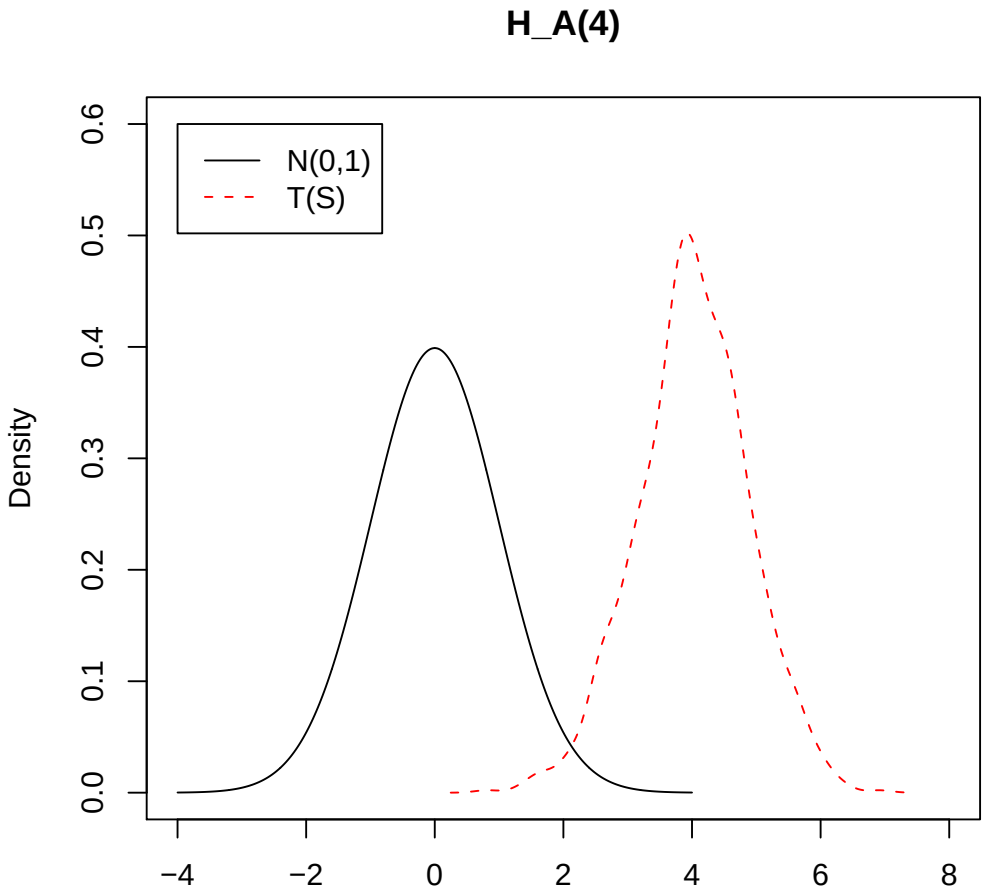

Figure 10: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(4)$, when p=1000, n=5, and $\rho = 0.1$.

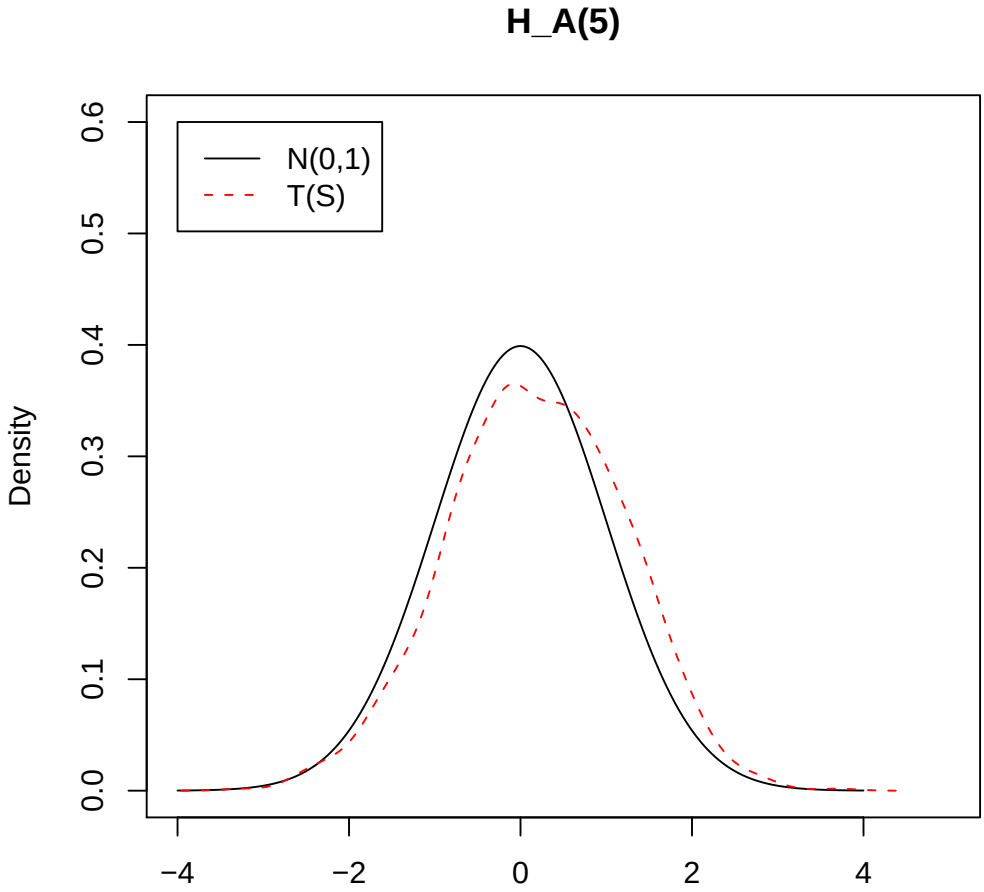

Figure 11: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(5)$, when p=1000, n=5, and $\rho = 0.1$.

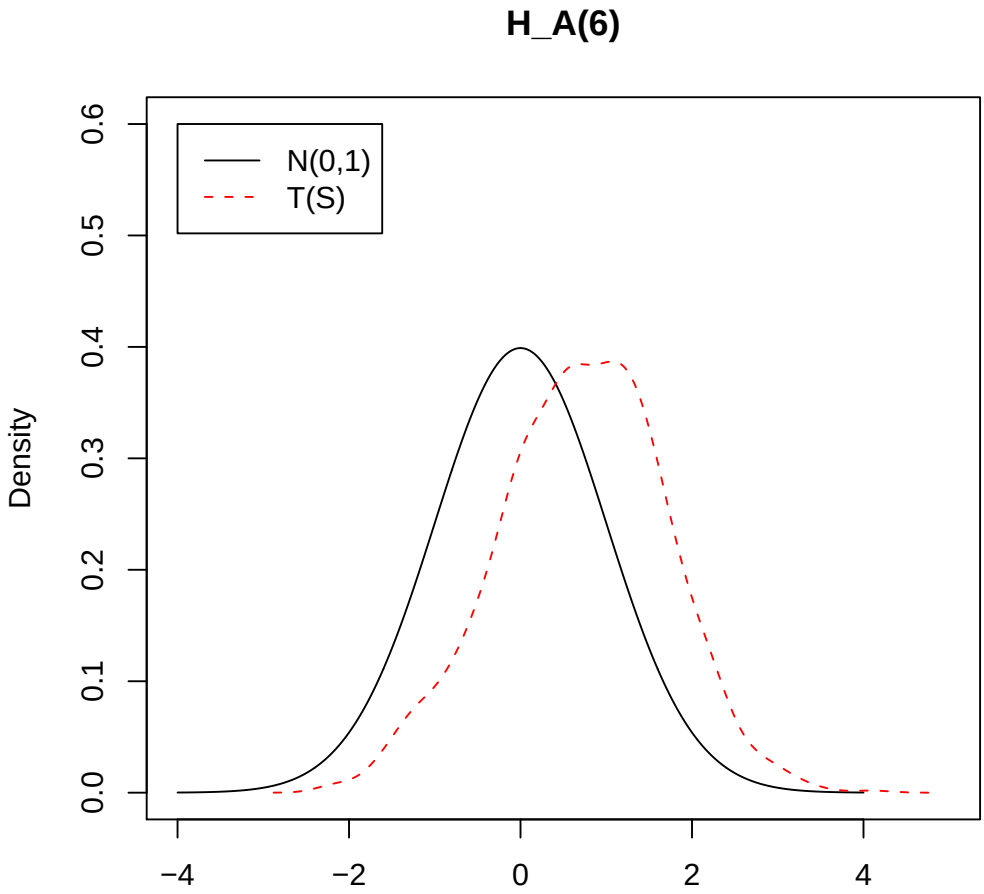

Figure 12: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(6)$, when p=1000, n=5, and $\rho = 0.1$.

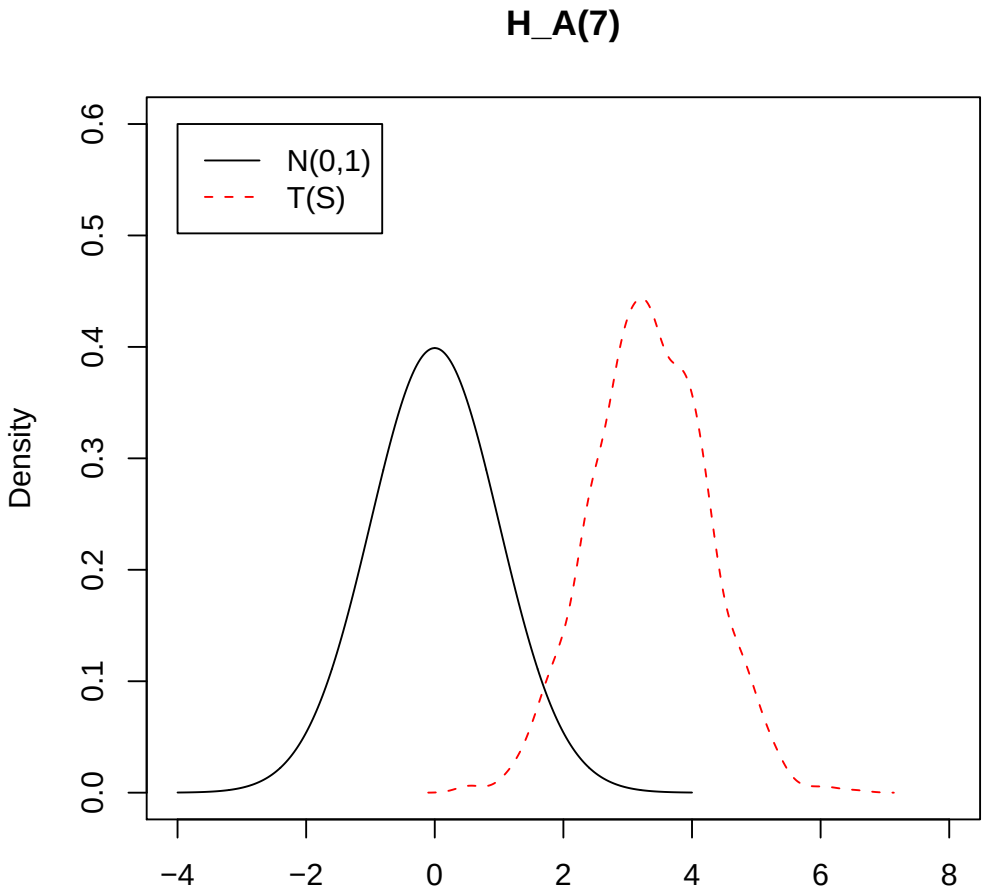

Figure 13: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(7)$, when p=1000, n=5, and $\rho = 0.1$.

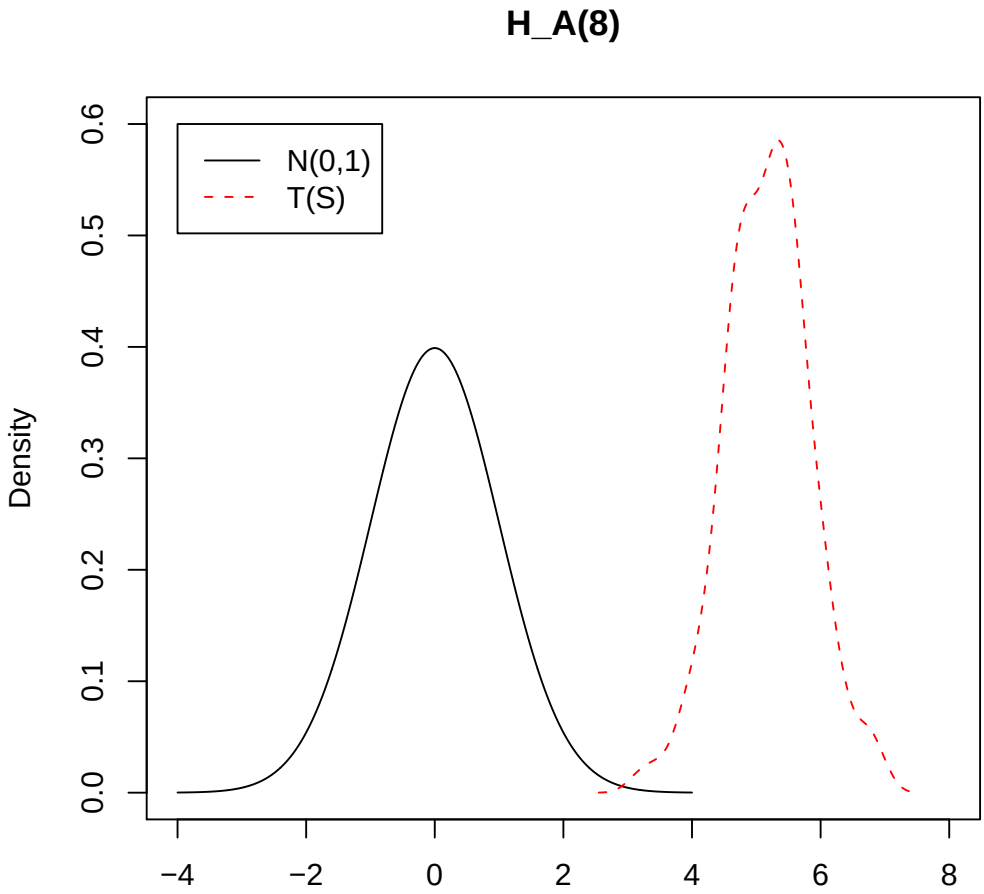

Figure 14: This graph plots a kernel estimate of the density of $T^{(S)}$ under $H_A(8)$, when p=1000, n=5, and $\rho = 0.1$.

3.4.2 Empirical Powers under Different Alternatives

In this section, we give a power table for each alternative from Table 6 to Table 13. Each empirical power is the percentage of cases in which the null hypothesis was rejected. We also create some figures for selected alternatives when $n = 3$, $\alpha = 0.05$. Please see Figure 15 to Figure 22. From the tables or figures, we can see that as p increases, the power increases accordingly. Also, the bigger n is, the greater the power, which agrees with one's expectations. It is also obvious that the bigger ρ yields the greater power. One would expect power to increase as ρ increases from 0 to $\frac{1}{2}$.

Take $H_A(1)$ as an example. Observing Table 6, we find

1. For $\rho = 0.1$, the powers are over 80% when $p \geq 1000$, and $n \geq 3$. When n is as few as 2, the powers are close to 100% if p is very large, say, 5000. If we increase the sample size from 3 to 5, then, when $p \geq 500$, the powers are over 90%.

2. For $\rho = 0.2$, the powers are all greater than those for $\rho = 0.1$. When $p = 1000$, the power is good even if $n = 2$. When $n = 3$, $p = 500$, the powers are close to 100%.

Note that our test has low power for $H_A(5)$ even when p is as large as 5000. This can be explained by Figure 5 as the alternative is so close to the null in location and scale. Nevertheless, it still shows the trend that the power increases as p increases.

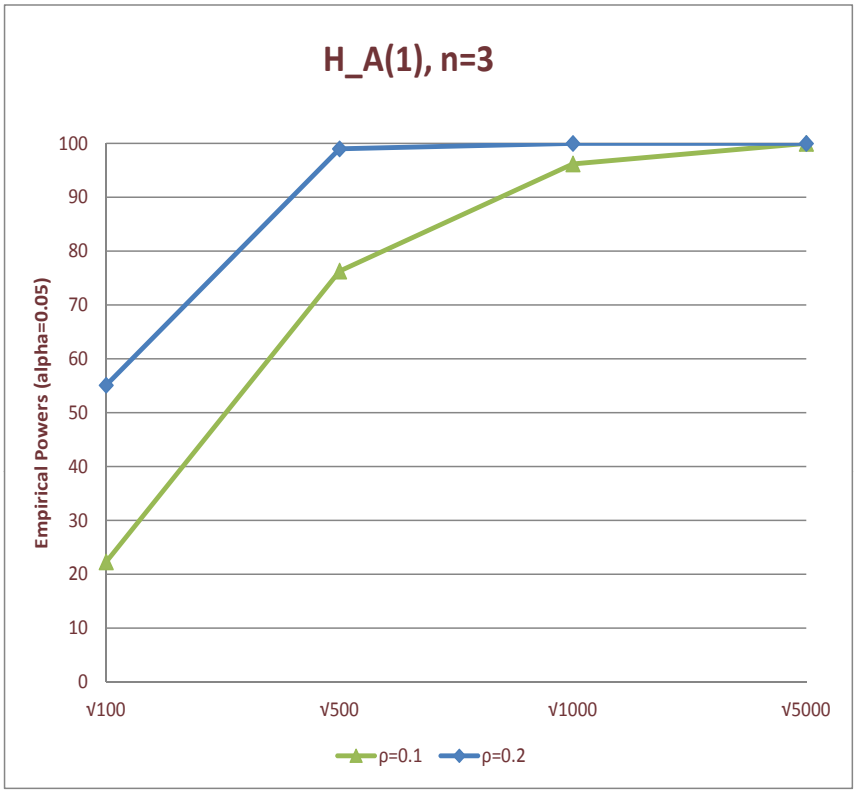

Figure 15: This graph compares the powers for $H_A(1)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

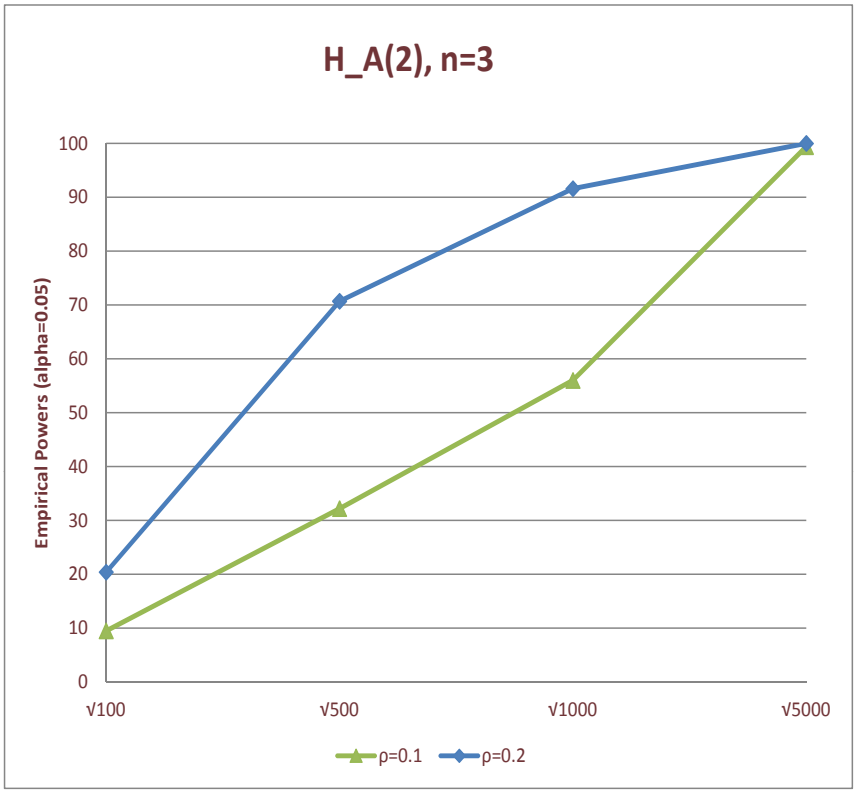

Figure 16: This graph compares the powers for $H_A(2)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

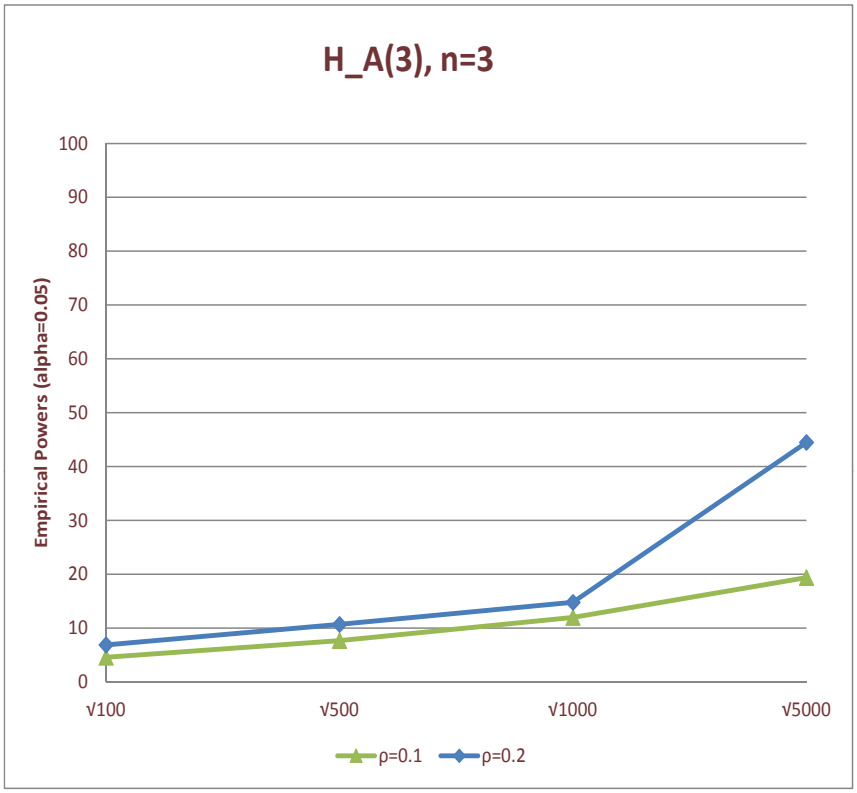

Figure 17: This graph compares the powers for $H_A(3)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

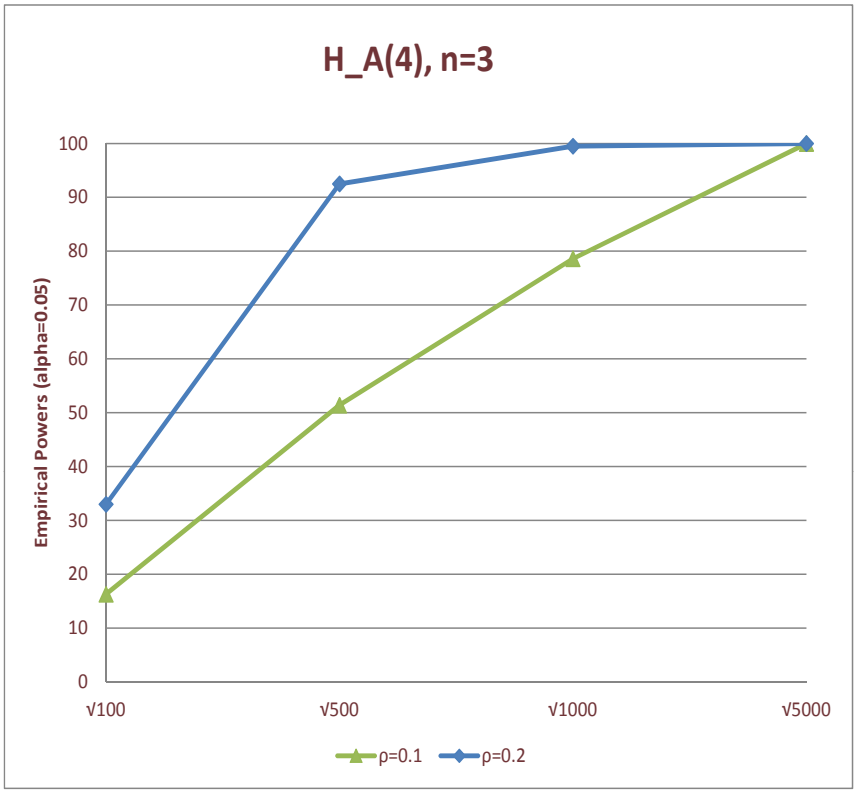

Figure 18: This graph compares the powers for $H_A(4)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

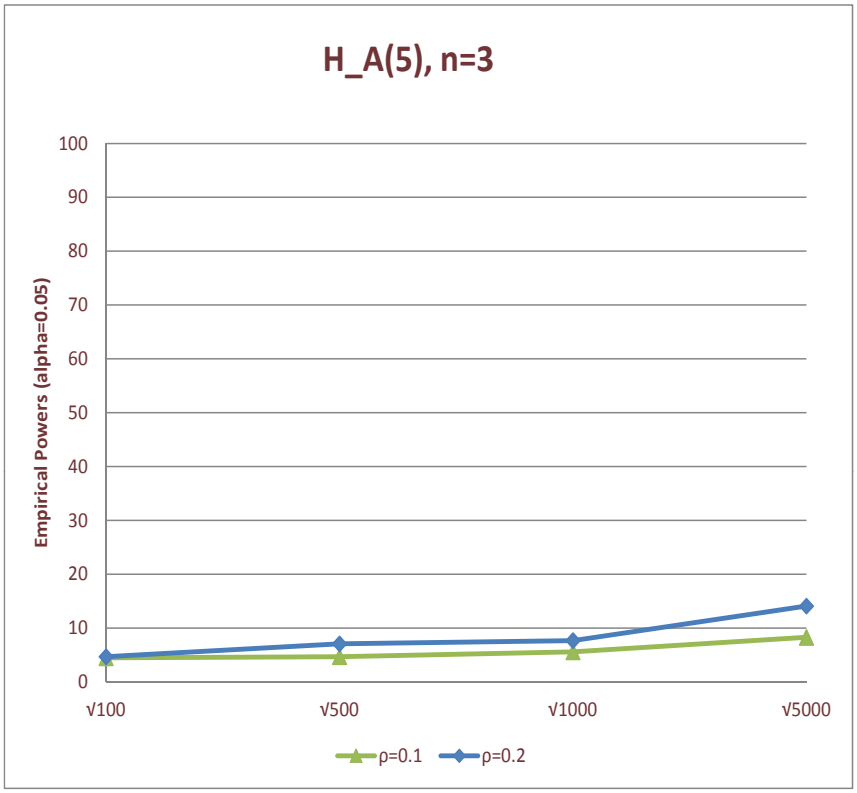

Figure 19: This graph compares the powers for $H_A(5)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

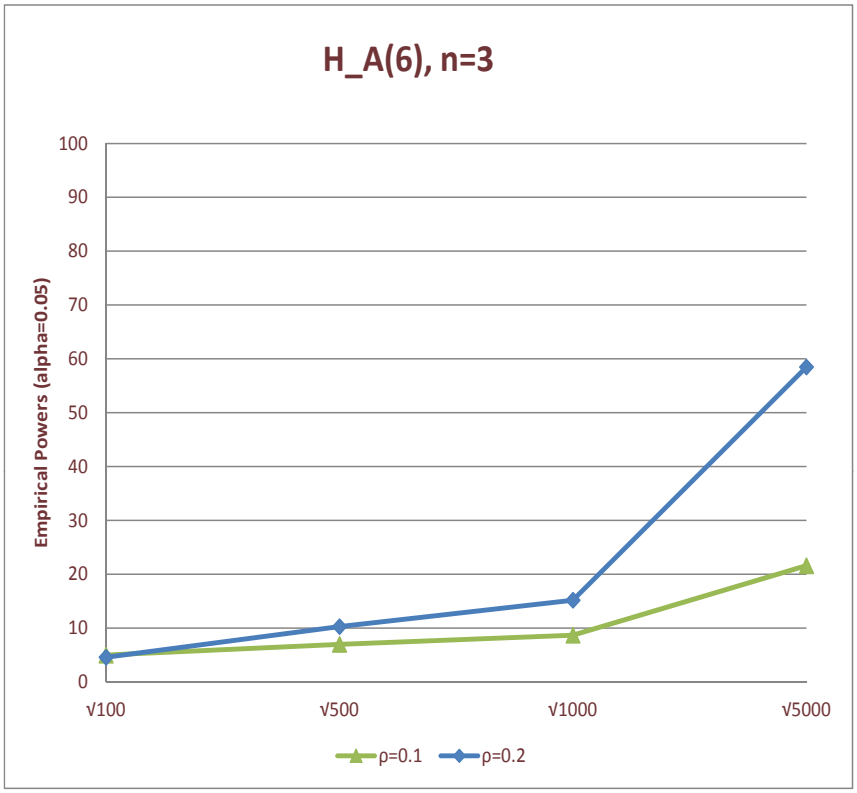

Figure 20: This graph compares the powers for $H_A(6)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

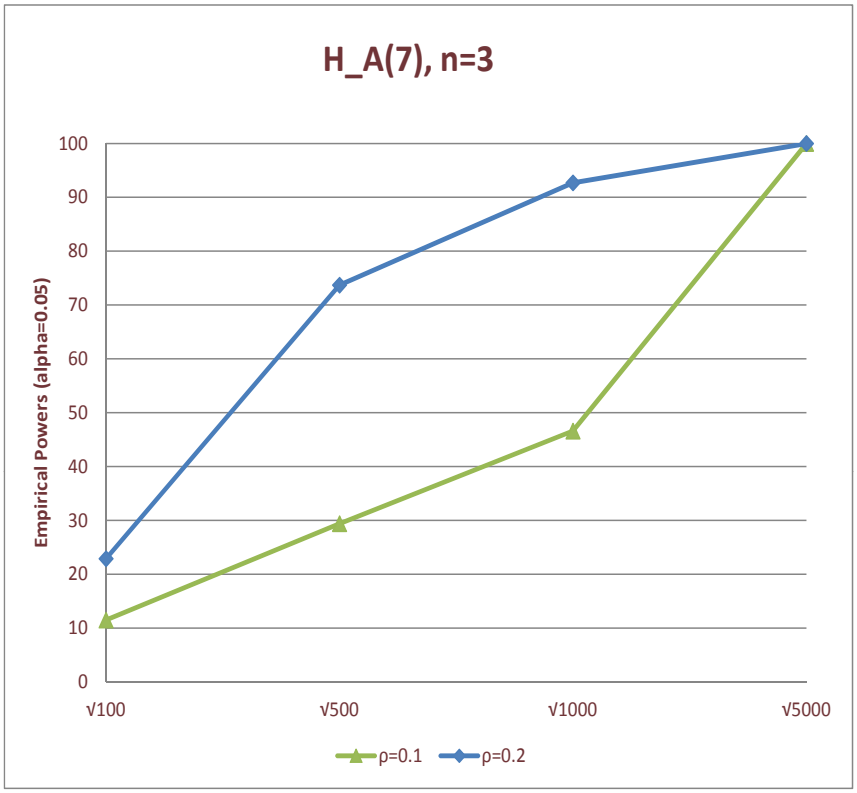

Figure 21: This graph compares the powers for $H_A(7)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

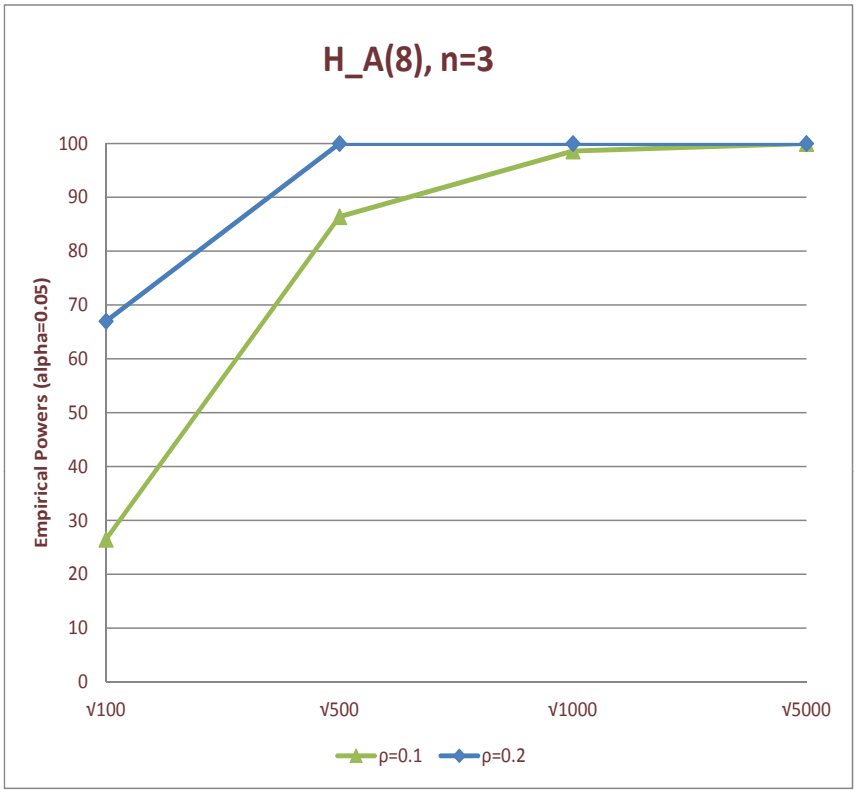

Figure 22: This graph compares the powers for $H_A(8)$, when $\rho = 0.1$ and 0.2. The x-axis is $\sqrt{p}$, where p is the number of data sets. Sample size n=3. $\alpha = 0.05$.

Table 6: Powers (%) for $H_A(1)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	3.7	12.7	24.1	9.9	29.2	44.5
	500	20.7	43.0	56.1	57.4	84.4	92.1
	1000	39.9	66.0	80.7	90.5	97.9	99.2
	5000	99.3	99.8	100	100	100	100
3	100	5.6	22.3	37.9	23.6	55.1	71.2
	500	46.5	76.3	85.1	95.5	99.0	99.8
	1000	84.4	96.2	98.6	100	100	100
	5000	100	100	100	100	100	100
5	100	15.0	45.4	64.2	62.4	89.3	95.5
	500	94.2	99.1	99.9	100	100	100
	1000	99.9	100	100	100	100	100
	5000	100	100	100	100	100	100
10	100	50.9	86.4	95.7	99.6	100	100
	500	100	100	100	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100

Table 7: Powers (%) for $H_A(2)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	1.5	8.6	14.7	3.0	12.6	22.8
	500	4.7	17.0	27.5	13.5	34.8	48.1
	1000	7.7	26.1	40.1	27.4	51.6	65.5
	5000	49.2	76.1	86.6	93.9	98.2	99.5
3	100	2.2	9.5	19.8	5.1	20.4	34.8
	500	11.0	32.2	47.9	42.2	70.7	81.5
	1000	28.4	56.0	69.3	74.1	91.6	96.0
	5000	95.4	99.4	100	100	100	100
5	100	5.3	21.6	37.0	17.8	46.5	62.6
	500	38.3	68.6	81.8	93.1	98.7	99.5
	1000	76.0	93.5	96.2	100	100	100
	5000	100	100	100	100	100	100
10	100	17.5	51.3	69.1	74.3	94.1	97.5
	500	96.1	99.6	99.8	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100

Table 8: Powers (%) for $H_A(3)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	1.1	5.0	10.3	1.2	4.9	9.5
	500	1.2	5.2	12.4	1.6	7.7	15.3
	1000	2.0	7.6	15.2	2.0	7.6	15.2
	5000	3.1	13.8	23.1	4.5	18.3	35.5
3	100	1.2	4.6	11.1	1.0	6.9	13.7
	500	2.3	7.7	14.5	2.7	10.7	17.8
	1000	2.2	12.0	19.8	3.6	14.8	25.8
	5000	4.9	19.4	35.5	18.0	44.5	58.6
5	100	1.0	6.8	12.7	1.3	10.6	19.6
	500	4.1	12.8	22.9	7.4	26.0	37.2
	1000	5.2	19.0	30.8	17.4	40.2	53.4
	5000	30.5	59.4	70.4	89.1	97.3	98.2
10	100	1.6	12.0	22.6	6.8	26.8	41.9
	500	16.1	40.5	55.2	53.8	79.8	89.5
	1000	37.2	64.9	77.7	90.6	97.2	99.2
	5000	98.9	100	100	100	100	100

Table 9: Powers (%) for $H_A(4)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	1.9	8.9	18.6	4.9	19.1	33.5
	500	9.6	29.1	41.7	30.7	58.1	71.9
	1000	17.3	34.0	56.0	63.2	83.8	90.9
	5000	86.0	96.0	98.4	100	100	100
3	100	3.7	16.3	28.0	10.0	33.0	49.6
	500	26.3	51.4	66.2	73.9	92.5	95.8
	1000	54.7	78.6	90.6	97.4	99.5	99.8
	5000	99.9	100	100	100	100	100
5	100	10.0	33.2	50.2	39.4	71.7	86.4
	500	70.0	90.4	96.2	99.4	100	100
	1000	97.4	99.5	99.9	100	100	100
	5000	100	100	100	100	100	100
10	100	37.9	75.1	88.3	92.4	99.6	99.7
	500	99.9	100	100	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100

Table 10: Powers (%) for $H_A(5)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	0.4	4.1	8.6	0.7	3.9	9.8
	500	1.0	4.9	9.6	0.8	5.9	11.4
	1000	1.2	5.8	10.9	1.1	5.9	11.4
	5000	1.6	7.3	14.4	2.4	9.2	17.4
3	100	0.4	4.5	10.0	0.6	4.7	10.6
	500	0.9	4.7	10.2	1.5	7.1	13.7
	1000	0.9	5.6	12.8	1.6	7.7	13.9
	5000	2.4	8.3	16.2	4.2	14.1	23.8
5	100	0.5	4.1	9.6	0.8	4.7	10.0
	500	0.9	5.8	11.7	2.8	10.5	18.1
	1000	1.2	7.1	15.1	3.4	11.5	19.8
	5000	2.4	15.8	25.3	7.2	25.9	37.7
10	100	0.8	5.8	11.5	0.8	5.8	13.8
	500	2.4	8.1	16.1	4.3	14.3	26.1
	1000	2.8	12.2	23.6	6.9	22.7	36.9
	5000	12.2	30.9	45.8	39.7	68.0	80.2

Table 11: Powers (%) for $H_A(6)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	0.7	4.1	8.4	0.9	5.2	9.6
	500	1.1	5.4	11.4	1.0	6.7	13.8
	1000	0.7	5.5	11.0	2.0	9.7	17.7
	5000	1.5	8.1	17.7	9.7	23.9	37.2
3	100	0.7	5.0	12.0	1.0	4.6	9.7
	500	1.4	7.0	14.9	2.0	10.3	20.8
	1000	1.8	8.7	16.2	4.2	15.2	26.8
	5000	6.6	21.6	28.1	29.9	58.5	68.9
5	100	1.2	6.8	14.4	1.6	9.4	17.6
	500	2.4	11.9	22.4	8.4	26.1	39.8
	1000	3.5	16.2	28.5	16.6	42.5	57.4
	5000	19.0	44.3	65.3	80.9	95.6	98.5
10	100	1.7	12.7	22.0	7.0	24.1	38.9
	500	13.1	36.6	50.7	51.9	79.8	89.2
	1000	31.1	56.0	70.6	87.4	96.1	98.3
	5000	100	100	100	100	100	100

Table 12: Powers (%) for $H_A(7)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	2.5	9.5	14.6	4.6	14.1	23.8
	500	5.0	17.4	25.7	14.3	31.0	44.8
	1000	6.3	19.3	29.6	22.5	47.2	60.0
	5000	19.7	55.3	67.1	93.2	100	100
3	100	3.1	11.5	21.8	8.0	22.9	38.8
	500	11.0	29.4	42.6	47.0	73.7	83.6
	1000	20.6	46.6	62.6	77.5	92.7	95.7
	5000	92.4	100	100	100	100	100
5	100	7.4	26.3	39.4	37.2	66.7	78.9
	500	51.4	78.4	87.8	99.5	99.9	100
	1000	88.0	97.3	99.1	100	100	100
	5000	100	100	100	100	100	100
10	100	44.4	79.5	92.0	99.6	100	100
	500	100	100	100	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100

Table 13: Powers (%) for $H_A(8)$.

		$\rho = 0.1$			$\rho = 0.2$		
n	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
2	100	3.3	14.9	26.3	13.3	39.7	55.4
	500	21.7	50.1	65.1	79.9	94.9	97.9
	1000	51.9	76.1	87.1	98.2	100	100
	5000	100	100	100	100	100	100
3	100	6.3	26.5	43.1	34.0	67.0	79.2
	500	62.5	86.4	92.9	100	100	100
	1000	91.1	98.6	99.6	100	100	100
	5000	100	100	100	100	100	100
5	100	15.1	48.2	65.8	68.7	92.8	98.0
	500	96.4	99.9	100	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100
10	100	30.5	79.0	91.7	93.7	99.6	99.9
	500	100	100	100	100	100	100
	1000	100	100	100	100	100	100
	5000	100	100	100	100	100	100

3.5 Benchmark Comparison

In this section, we compare our test with the classic one-way ANOVA F test and the non-parametric Kruskal-Wallis (K-W) test, when the data sets only differ by location. According to 5000 simulations, the F test has probabilities of type I errors close to the nominal significance levels. The K-W test has the empirical probability of type I errors lower than the nominal levels, although it seems that as n increases, the true significance level is getting closer and closer to the nominal levels. However, in order to compare the powers among these three tests, we simulate the critical values for the K-W test and use them to get the powers in Table 14. For comparison purpose, we only choose $H_A(1)$ with $p = 100, 1000$, and $\rho = 0.1$.

Table 14 shows that: 1) The powers for all three tests increase as p increases from 100 to 1000. For a fixed p, power increases as n increases; 2) When $p = 1000$, the power of our test gets closer to that of the F test or K-W test, which means when p is a large number, our test is comparable to either the F test or K-W test. For $\alpha = 0.05$, we have plotted the comparison in Figure 23.

Table 14: Empirical powers (%) for tests: 1) $T^{(S)}$; 2) F test; and 3) Kruskal-Wallis (K-W) test under $H_A(1)$, and $\rho = 0.1$. The K-W test uses the simulated critical values.

			Power (%)		
p	n	Test	α = 0.01	α = 0.05	α = 0.10
100	2	$T^{(S)}$	3.7	12.7	24.1
		F	7.1	21.4	33.9
		K-W	6.2	17.5	30.4
	3	$T^{(S)}$	5.6	22.3	37.9
		F	17.6	40.6	54.9
		K-W	15.5	36.3	49.5
	5	$T^{(S)}$	15.0	45.4	64.2
		F	54.7	78.6	87.3
		K-W	45.8	71.7	81.2
	10	$T^{(S)}$	59.9	86.4	95.7
		F	98.7	99.8	99.9
		K-W	96.9	99.4	99.8
1000	2	$T^{(S)}$	39.9	66.0	80.7
		F	62.1	83.7	90.9
		K-W	47.5	73.8	84.2
	3	$T^{(S)}$	84.4	96.2	98.6
		F	97.9	99.6	99.9
		K-W	94.7	98.9	99.7
	5	$T^{(S)}$	99.9	100	100
		F	100	100	100
		K-W	100	100	100
	10	$T^{(S)}$	100	100	100
		F	100	100	100
		K-W	100	100	100

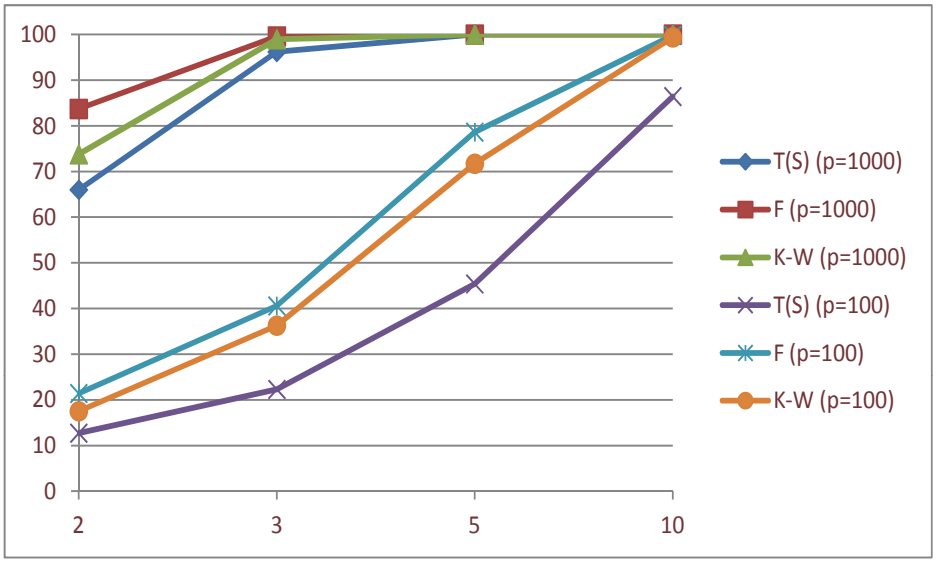

Figure 23: Comparison of powers for tests: 1) $T^{(S)}$; 2) F test; and 3) Kruskal-Wallis (K-W) test under $H_A(1)$, and $\rho = 0.1$. $\alpha = 0.05$. The x-axis indexes the sample size n. The y-axis indexes the power in percent. The K-W test uses the simulated critical values.

CHAPTER IV

REAL DATA ANALYSES

4.1 Steps for Conducting Test

In practice, the following steps are followed in conducting our test:

1. Calculate the over-smoothed bandwidth: h_{os}.
2. Compute test statistic: $T^{(S)}$.
3. Choose a critical value: Z_α, which is the upper α quantile of the standard normal distribution.
4. Make a decision by comparing $T^{(S)}$ with the critical value.

In the following sections, we first give a brief background of some microarray data, and then apply our test to those data.

4.2 Background of the Rat Data

Our real data are microarray data, which we refer to as the "rat" data. These were collected by Robert Chapkin and coworkers at Texas A&M University. As stated in Davidson, Nguyen, Hokanson, Callaway, Isett, Turner, Dougherty, Wang, Lupton, Carroll, and Chapkin (2004), they used Codelink DNA microarrays containing about 9000 genes to help decipher the global changes in colonocyte gene expression profiles in carcinogen-injected Sprague Dawley rats. However, the data we analyze are the same as those used by Hart and Cañette (2011). Hart and Cañette (2011) applied a rank test to test whether an $LSRE$ model is more appropriate than an LRE model. In other words, they tried to detect if

there exist scale differences from one gene to another. It is concluded from their study that differences in scale exist.

Below are some descriptions of the rat data from Hart and Cañette (2011):

The data are an 8038×5 matrix, i.e., $p = 8038, n = 5$. Each row represents the data from one gene. Each column represents 8038 genes from one rat. There are 5 rats, with 8038 genes for each one.

The five rats from which these data were collected were all subjected to the same treatment. The original data are $Y_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5$, where i indexes genes, j indexes different rats, and Y_{ij} is the logarithm of the expression level for gene i and rat j.

The following model for the data were assumed:

$$Y_{ij} = \mu_i + R_j + \epsilon_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5,$$

where R_j represents a rat effect, μ_i a gene effect, and ϵ_{ij} measurement error.

After estimating rat effects by computing the mean of all data for each rat, Hart and Cañette (2011) defined the centered data set after removing the rat effects as follows:

$$X_{ij} = Y_{ij} - \frac{1}{8038} \sum_{i=1}^{8038} Y_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5,$$

4.3 Test Applied to Centered Data

We want to test if all the data sets $(X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}), i = 1, \ldots, p$, are from the same distribution.

Figure 24 shows scatter plots for each rat. Hart and Cañette (2011) argue that there is little evidence of correlation from gene to gene. We calculate the sample standard deviation within each small data, and find the maximum to be 4.1. The pooled standard deviation is $s_{pool} = 0.266$. Therefore, the over-smoothed bandwidth for the centered rat data is calculated to be: $h_{os} = 0.221$.

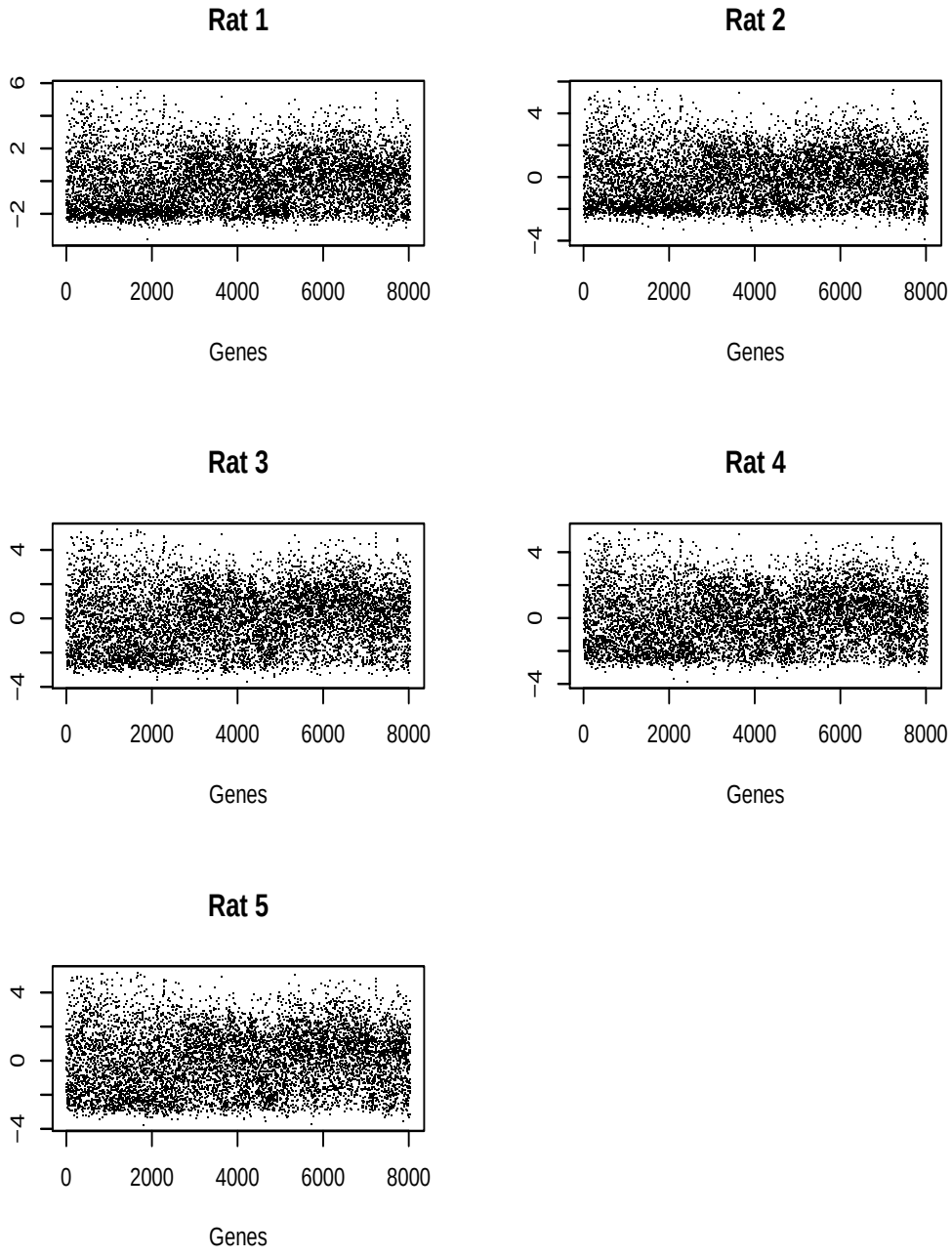

Figure 24: These scatter plots are from the centered data for each of five rats, with the number of genes equalling 8038 for each rat.

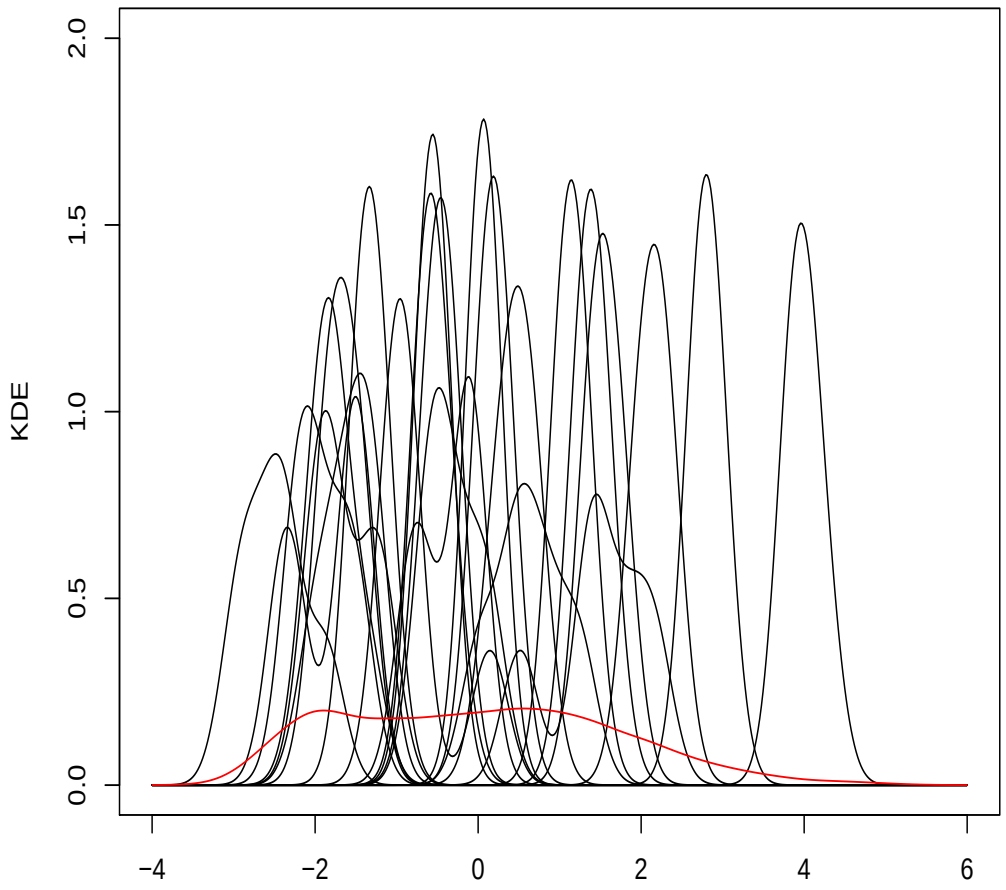

Figure 25: This graph plots kernel density estimates with the over-smoothed bandwidth calculated from the centered rat data (bandwidth = 0.221). The black curves are from the first 25 genes. The red curve is the overall kernel density estimate from all the small data sets.

Figure 25 plots the kernel density estimates for the first 25 small data sets (genes) and the overall kernel density estimate in red. If we just eyeball the curves, they do appear to be different, especially in terms of location. Therefore, intuitively, we should reject the null hypothesis in this case.

Using the R-function **TS** in Appendix III for $T^{(S)}$, we have the following results as shown in Table 15:

Table 15: Results of applying our test to the centered rat data.

h	$T_p^{(S)}$	σ_S	$T^{(S)}$	S_W	S_B
0.221	0.777	0.2486	280.22	0.941	0.164

Comparing S_W with S_B, we notice there is a large discrepancy between these two numbers (0.941 vs. 0.164). This discrepancy indicates that the distributions within genes may be significantly different from the overall distribution. Comparing $T^{(S)}$ with the critical values from the standard normal distribution, we reject H_0 and conclude that there is significant evidence to show that not all the centered data are from a common distribution. This conclusion agrees with the intuitive result from visually looking at the curves in Figure 25.

In case we have a very large p, we may conduct the test by applying it to sub-samples. In this case, we have p=8038. It works if we sample 1000 data sets from the total 8038 small data sets. Note that when doing the sub-sampling, we need to make sure we sample the entire i-th small data set if i is selected. We have tried two random sub-samples without replacement from the rat data and get the following statistics in Table 16, which lead us to the same conclusion as before.

Table 16: Results of applying our test to the centered rat data when sampling 1000 small data sets from the entire data set.

Sub-sample	h	$T_p^{(S)}$	σ_S	$T^{(S)}$	S_W	S_B
1	0.22	0.7732	0.264	92.53	0.9385	0.1653
2	0.194	0.8567	0.281	96.33	1.018	0.1613

4.4 Test Applied to Transformed Data

In microarray analyses, it is often reasonable to assume that these p data sets have some degree of commonality, as exemplified by the following model:

$$X_{ij} = \mu_i + Z_{ij}, i = 1, \ldots, p, j = 1, \ldots, n, \tag{4.1}$$

where X_{ij} are observed real-valued data, and $Z_{ij}, i = 1, \ldots, p, j = 1, \ldots, n$, are unobserved errors.

Following Hart and Cañette (2011), we make the following assumptions:

A1. The μ_i, $i = 1, \ldots, p$, are independent, identically distributed and unknown parameters.

A2. The $Z_{ij}, i = 1, \ldots, p, j = 1, \ldots, n$, are mutually independent, and $Z_{i1}, \ldots, Z_{in}$ have cumulative distribution function (CDF) F_i, $i = 1, \ldots, p$. Each Z_{ij} has mean 0.

A3. The parameters μ_i, $i = 1, \ldots, p$, are independent of $Z_{ij}, i = 1, \ldots, p, j = 1, \ldots, n$.

In model (4.1), if $F_1 = \ldots = F_p$, then, it is called a location random effects (LRE) model since the distributions for the p data sets differ only with respect to location μ_i.

In the following, we want to test equality of the distributions $F_1, \ldots, Fp$ in model (4.1).

We may write that, under the null, to a good approximation, the X_{ij}'s have the following LRE model:

$$X_{ij} = \mu_i + Z_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5,$$

where $Z_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5$, are unobserved errors with a common distribution F, such that $F_1 = \cdots = F_p = F$.

To apply our test, we need to transform the data at the first step in order to remove $\mu_1, \ldots, \mu_p$. There are two ways to remove these parameters. One is called "residuals", and the other "differences". We will illustrate these two different procedures separately in Section 4.4.1 and Section 4.4.2.

An alternative to the LRE model is the location scale random effect ($LSRE$) model, which has the following form:

$$X_{ij} = \mu_i + \sigma_i Z_{ij}, i = 1, \ldots, 8038, j = 1, \ldots, 5,$$

where the only difference from (4.2) is that σ_i's are allowed to be different from one gene to another.

Hart and Cañette (2011) devised a test of the LRE versus the $LSRE$ model and applied it to the rat data. They concluded that the LRE model could be rejected in favor of the $LSRE$ model. Since Hart and Cañette (2011) rejected $F_1 = \cdots = F_p$ in favor of scale differences among the F_i's, it will be interesting to see if our test rejects $H_0 : F_1 = \cdots = F_p$.

4.4.1 Transformation of "Residuals"

Our method will be applied to the "residuals", $\delta_{ij} \equiv X_{ij} - \bar{X}_i^{(j)}$, where

$$\bar{X}_i^{(j)} = \frac{1}{n-1} \sum_{k=1, k \neq j}^{n} X_{ik},\ i = 1, \ldots, p,\ j = 1, \ldots, n.$$

Note that $\delta_{ij} = X_{ij} - \bar{X}_i^{(j)} = Z_{ij} - \bar{Z}_i^{(j)}$ under the null, where $\bar{Z}_i^{(j)} = \frac{1}{n-1}\sum_{k=1,k\neq j}^{n} Z_{ik}$, $i = 1, \ldots, p$, $j = 1, \ldots, n$.

One may question the identifiability of the distributions of Z_{ij}, denoted as F_i, $i = 1, \ldots, p$. The good news is that under the LRE Model, it has been shown by Hart and Cañette (2011) that when $n \geq 3$, the distribution F of Z_{ij} is identifiable from that of δ_{ij}, as long as the characteristic function of F does not vanish throughout an interval.

Therefore, suppose that δ_{ij}'s have the distribution functions G_i, $i = 1, \ldots, p$. Testing equality of F_i's is thus equivalent to testing equality of G_i's, $i = 1, \ldots, p$. In the following, we are actually testing $H_0 : G_1 = G_2 = \cdots = G_p = G$, where G denotes the common cumulative distribution of δ_{ij}'s.

The pooled standard deviation for the "residuals" is about 0.33, and the over-smoothed bandwidth is calculated to be 0.276.

Because the transformed rat data no longer have independence within each small data set, the mean of $T_p^{(S)}$ under the null does not necessarily equal zero. Therefore, we can not apply the standardized test statistic $T^{(S)}$ and utilize the critical values from the standard normal distribution any more. However, Hart and Cañette (2011) estimated the error distribution with their methodology. Therefore, by using their estimated error distribution, we can still conduct our test by comparing $T_p^{(S)}$ with the critical values from a bootstrap procedure. We apply the bootstrap as follows: Firstly, we generate B different 8038×5 matrices from the estimated error distribution. Then,we calculate the transformed data $\delta_{ij}^{(B)}$ as we have done for the centered data. Finally, we compute $T_p^{(S)}$ B times and use the 95th percentile as the critical value for a level 5% test. This bootstrap procedure gives us the following critical value: $T_{0.05,boot} = -0.00475$.

Figure 26 plots the distribution of $T_p^{(S)}$ from the bootstrap procedure when B=1000. The results in Table 17, and also Figure 26, show that we should reject the null hypothesis.

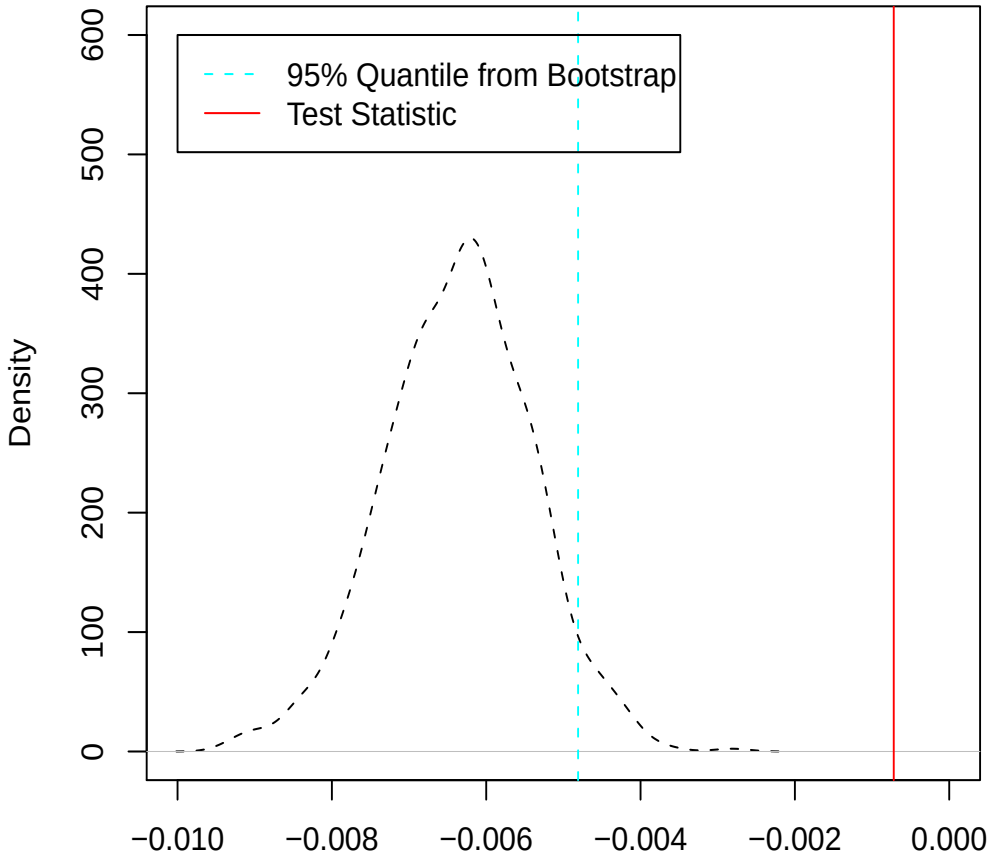

Figure 26: A kernel estimate of the density of the unstandardized test statistics, $T_p^{(S)}$. The number of bootstrap replications was 1000. The red line indicates the value of $T_p^{(S)}$ from the rat data, and the light blue dashed line indicates the 95% quantile of the distribution from bootstrapping.

Table 17: Results of applying our test to the transformed rat data of "residuals".

h	$T_p^{(S)}$	S_W	S_B	$T_{0.05,boot}$
0.276	-0.0007189	0.75276	0.75348	-0.00475

4.4.2 Transformation of "Differences"

However, by conducting the test in section 4.4.1 , we need to utilize knowledge of the underlying density of the errors, which is obtained from Hart and Cañette (2011). What if we have no such knowledge? Generally, another transformation can be considered, which is calculating the "differences", d_{ijk}. Specifically, $d_{ijk} = z_{ij} - z_{ik}$, $i = 1, \ldots, p$, $j, k = 1, \ldots, n$, $j \neq k$. Suppose that d_{ijk} has the true density G_i, $i = 1, \ldots, p$. We are testing the null hypothesis of the equality of all the distributions $G_1, \ldots, G_p$.

The advantage of this transformation is that the differences will keep the independence property within each small data set, and we can apply the standardized test statistic $T^{(S)}$ directly. The disadvantages are in two aspects: identifiability and power.

Clearly, $G_1 \neq G_2 \Rightarrow F_1 \neq F_2$, and hence rejection of $H_0 : G_1 = \cdots = G_p$ allows us to reject $H_0 : F_1 = \cdots = F_p$. The problem is that $G_1 = G_2$ does not necessarily imply that $F_1 = F_2$, and so "acceptance" of H_0 does not necessarily entail acceptance of $F_1 = \cdots = F_p$.

Another issue is the power of the test. Previously, the transformation of "residuals" doesn't change the sample size, but using independent "differences" will. Take the rat data as an example. Originally, we have 5 observations for each gene, i.e., n=5. But only 2 differences are independent. Therefore, the transformation of "differences" hurts the power of the test to some extent, although not so much if p is large enough. However, this transformation does require the sample size to be at least four in order to conduct the test.

For the rat data, we calculate the differences between two rats by randomly selecting four rats without replacement for each gene. We then get a new data matrix of 8038 rows and 2 columns. Applying our test statistic using the R function **TS** in Appendix III, we have the results for one set of "differences" in rat data "differences" set 1 in Table 18, and a figure indicating the scale differences in rat data "differences" set 1 is shown in Figure 27.

Table 18: Results of applying our test to the transformed rat data of "differences". Rat data "differences" set 1 is used.

h	$T_p^{(S)}$	σ_S	$T^{(S)}$	S_W	S_B
0.37	0.0179	0.1541	10.446	0.5989	0.581

Observing Figure 27, we can see there exist some scale differences for the first 25 curves. Comparing $T^{(S)}$ with the standard normal critical values, we feel confident to reject the null and conclude that there are differences between the error distributions of different genes. To confirm our conclusion, we also calculate several other sets of "differences". They all lead to quite similar test statistics. We summarize these numbers in Table 19.

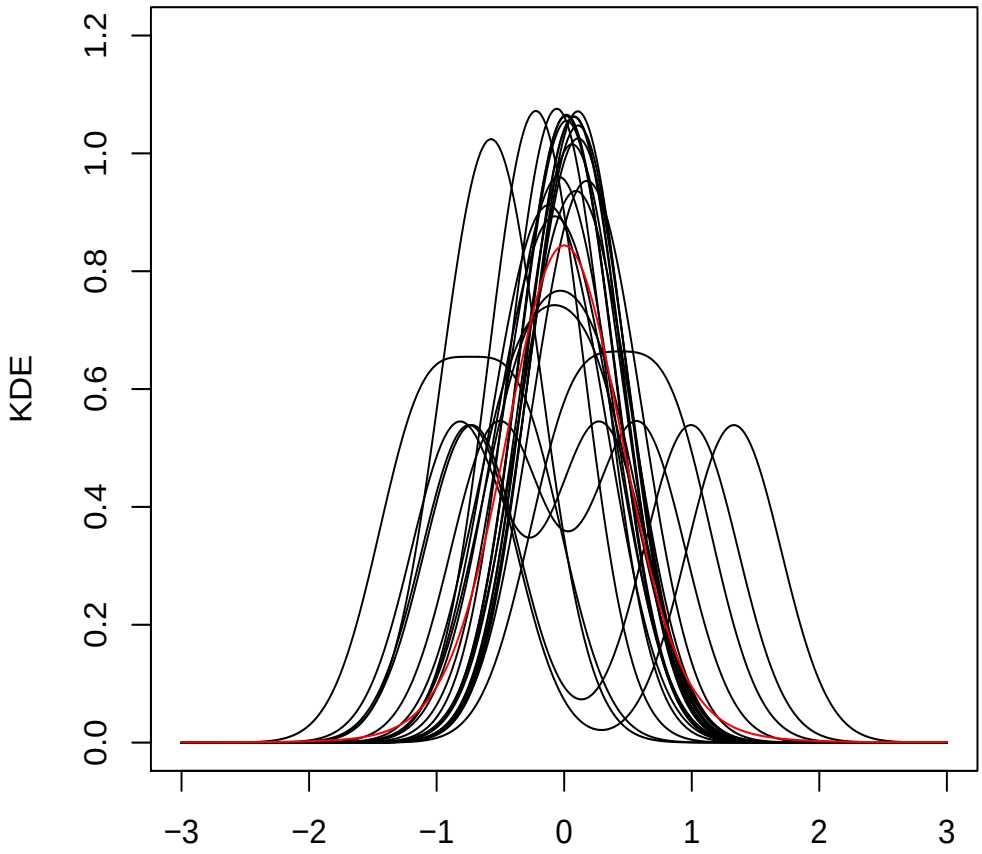

Figure 27: This graph plots kernel density estimates with the over-smoothed bandwidth calculated from the transformed rat data "differences" set 1 (bandwidth = 0.37). The black curves are from the first 25 genes. The red curve is the overall kernel density estimate from all the small data sets.

Table 19: Results of applying our test to the transformed rat data of "differences". Rat data "differences" sets 2, 3, and 4 are used.

Rat data "differences"	h	$T_p^{(S)}$	σ_S	$T^{(S)}$	S_W	S_B
Set 2	0.37	0.0153	0.1503	9.1076	0.5960	0.5807
Set 3	0.366	0.0173	0.1534	10.11	0.6027	0.5854
Set 4	0. 39	0.0152	0.1399	9.741	0.5774	0.5622

CHAPTER V

SUMMARY AND FUTURE RESEARCH

5.1 Summary

In this dissertation, we study the k-sample problem when k is large and n small. Actually, we are testing equality of a large number (p) of distributions, with only a small number of observations from each distribution. The null hypothesis is that all the small data sets are from the same distribution. We propose a test statistic, $T_p^{(S)}$, which is based on kernel density estimates and has expected value 0 under the null, i.e., $E(T_p^{(S)}|H_0)$=0. When p goes to infinity, but sample size (n) is bounded, we obtain the asymptotic distribution of $T_p^{(S)}$ to be normal. We estimate the variance of $T_p^{(S)}$ and get the standardized version of the test statistic, $T^{(S)}$, which has the limiting distribution of $N(0,1)$. Our test can detect any sort of differences among the distributions so long as the number of data sets that have the same distributions is not too large. It is shown that $T^{(S)}$ is invariant to a location and scale transformation of the data. Simulation studies show that our test has good power against a variety of alternatives.

Our methodology has applications to microarray analyses, where one usually encounters thousands of genes, but only a few replications. We apply our method to a real "rat" data set, which has 8038 genes but only 5 mice. The real data analyses show that our test finds differences between gene distributions that are not due simply to location.

5.2 Future Research

In this dissertation, we assume the sample size n is the same for each small data set, which might not be true in real life especially when missing data exist. Part of our future research is to extend our methodology to the case when sample sizes are different. As for the band-

width selection in this research, we may further investigate if our over-smoothed bandwidth is within the range that produces great powers.

When the LRE model is assumed, our methodology is limited. We need to transform the data by either "residuals" or "differences" as in Section 4.4. Neither of them provides a perfect solution. In the future, we may seek another method to make the test work better under the LRE or $LSRE$ model. We may consider more complicated models, such as a model that allows both mean and standard deviation to vary from one small data set to another, even under the null hypothesis. In the case when all the small data sets are not independent, it is also important and interesting for future research.

We also wish to investigate diagnostic methods to identify which distributions are different from the majority in the case that we reject the null hypothesis.

Future work also includes considering a different type of statistic, such as a test statistic of likelihood ratio type.

REFERENCES

Anderson, N., Hall, P., and Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariante Analysis* **50**, 41–54.

Anderson, T. W. (1962). On the distribution of the two-sample cramer-von mises criterion. *Annals of Mathematical Statistics* **33**, 1148–1159.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association* **49**, 765769.

Cao, R. and Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *The Canadian Journal of Statistics* **34**, 61–67.

Corder, G. W. and Foreman, D. I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. New Jersey: Wiley.

Davidson, L. A., Nguyen, D. V., Hokanson, R. M., Callaway, E. S., Isett, R. B., Turner, N. D., Dougherty, E. R., Wang, N., Lupton, J. R., Carroll, R. J., and Chapkin, R. S. (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research* **64**, 6797–6804.

Dudoit, S., Shafer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

Epanechnikov, V. (1969). Nonparametric estimation of a multivariate probability density. *Theory of Probability and Applications* **14**, 153–158.

Eubank, R., Hart, J., and LaRiccia, V. (1993). Testing goodness of fit via nonparametric function estimation techniques. *Communications in Statistics* **22**, 3327–3354.

Hart, J. D. (1997). *Nonparametric Smoothing and lack-of-Fit Tests*. New York: Springer.

Hart, J. D. and Cañette, I. (2011). Nonparametric estimation of distributions in random effects models. *Journal of Computational and Graphical Statistics* **20**, 461–478.

Hartley, H. (1950). The use of range in analysis of variance. *Biometrika* **37**, 271280.

Iversen, G. R. and Norpoth, H. (1987). *Analysis of Variance*. Thousand Oaks, CA: Sage Publications, Inc.

Jimenez-Gamero, M., Albr-Fernandez, V., Munoz-Garcia, J., and Chalco-Cano, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions. *Computational Statistics and Data Analysis* **53**, 3957–3971.

Kiefer, J. (1959). k-sample analogues of the kolmogorov-smirnov, cramer-von mises test. *Annals of Mathematical Statistics* **30**, 420–447.

Kruskal, W. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, (260): 583–621.

Levene, H. (1960). Robust tests for equality of variances. *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, 278292.

Loader, C. R. (1999). Bandwidth selection: Classical or plug-in?. *The Annals of Statistics* **27**, 415–438.

Louani, D. (2000). Large deviation for l1-distance in kernel density estimation. *Journal of Statistical Planning and Inference* **90**, 177–1828.

Martínez-Camblor, P. and Uña-Álvarez, J. D. (2009). Non-parametric k-sample tests: Density functions vs distribution functions. *Computational Statistics and Data Analysis* **53**, 3344–3357.

Martínez-Camblor, P., Uña-Álvarez, J. D., and Corral, N. (2008). k-sample test based on the common area of kernel density estimator. *Journal of Statistical Planning and Inference* **138(12)**, 4006–4020.

Moore, D. S., McCabe, G., and Craig, B. (2007). *Introduction to the Practice of Statistics*. New York: Freeman W. H. & Company.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 10651076.

Plackett, R. (1983). Karl pearson and the chi-squared test. *International Statistical Review* **51(1)**, 59–72.

Rayner, J. C. and Best, D. J. (1989). *Smooth Tests of Goodness of Fit*. New York: Oxford University Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density functions. *Annals of Mathematical Statistics* **27**, 832–837.

Scholz, F. and Stephens, M. (1987). k-sample anderson-darling test. *Journal of American Statistics Association* **82**, 918–924.

Schucany, W. R. and Bankson, D. M. (1989). Small sample variance estiamtors for u-statistics. *Australian Journal of Statistics* **31**, 417–426.

Sen, P. K. (1960). On some convergence properties of u-stastistics. *Calcutta Statistical Association Bulletin* **10**, 1–18.

Serfling, R. J. (2002). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

Sheather, S. J. (2004). Density estimation. *Statistical Science* **19**, 588–597.

Silverman, B. W. (1978). Choosing window width when estimating a density. *Biometrika* **65**, 1–11.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*. Ames, IA: Iowa State University Press.

Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* **69**, 730–737.

Stephens, M. A. (1986). Tests based on edf statistics. *GoodnessofFit Techniques* **68**, 97–185.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association* **85**, 470–477.

Terrell, G. R. and Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association* **85**, 209–214.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Young, S. G. and Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics* **51**, 920–931.

APPENDIX I

PROOF OF ASYMPTOTIC DISTRIBUTION OF T_P

I.1 General Idea for Proof

$$T_p = \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}(\hat{f}_h(x|i) - \hat{f}_h^{(i)}(x))^2 dx \tag{I.1}$$

Under H_0, we have $E\hat{f}_h(x|i) = E\hat{f}_h^{(i)}(x) = E\hat{f}_h^{(i,j)}(x) \equiv f(x;h)$, where $\hat{f}_h^{(i,j)}(x)$ is a kernel density estimate computed from all the pooled data with both the i-th and the j-th data set excluded.

For the purpose of writing simplicity, the following notations are defined:

$$\begin{aligned}
\delta_i(x) &\equiv f(x;h) - \hat{f}_h(x|i) \\
\gamma_i(x) &\equiv f(x;h) - \hat{f}_h^{(i)}(x) \\
\gamma_{i,j}(x) &\equiv f(x;h) - \hat{f}_h^{(i,j)}(x).
\end{aligned}$$

Note that: $\delta_i(x), \delta_j(x), \gamma_{i,j}(x)$ are independent, and having the following properties under H_0:

$$E\delta_i(x) = E\gamma_i(x) = E\gamma_{i,j}(x) = 0 \tag{I.2}$$

$$E\delta_i^2(x) = \text{Var}\left(\hat{f}_h(x|i)\right) \tag{I.3}$$

$$E\gamma_i^2(x) = \text{Var}\left(\hat{f}_h^{(i)}(x)\right). \tag{I.4}$$

Thus,

$$\begin{aligned} T_p &= \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\left[(\hat{f}_h(x|i)-f(x;h))+(f(x;h)-\hat{f}_h^{(i)}(x))\right]^2 dx \\ &= \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\left[-\delta_i(x)+\gamma_i(x)\right]^2 dx \\ &= \frac{n}{p}\sum_{i=1}^{p}\left\{\int_{-\infty}^{\infty}\delta_i^2(x)dx+\int_{-\infty}^{\infty}\gamma_i^2(x)dx+2\int_{-\infty}^{\infty}-\delta_i(x)\gamma_i(x)dx\right\} \\ &\equiv A_p+B_p+C_p\ , \end{aligned}$$

where

$$\begin{aligned} A_p &\equiv \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\delta_i^2(x)dx \\ B_p &\equiv \frac{n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\gamma_i^2(x)dx \\ C_p &\equiv \frac{-2n}{p}\sum_{i=1}^{p}\int_{-\infty}^{\infty}\delta_i(x)\gamma_i(x)dx. \end{aligned}$$

Let $Y_i \equiv \int_{-\infty}^{\infty}\delta_i^2(x)dx$ $(i=1,\cdots,p)$, having $\mu \equiv EY_1 < \infty, \sigma^2 \equiv \text{Var}(Y_1)$. Then, by the Central Limit Theorem,

$$\frac{\frac{1}{p}\sum_{i=1}^{p}Y_i-\mu}{\sigma/\sqrt{p}} \xrightarrow{\mathcal{D}} N(0,1), \text{ as } p\to\infty.$$

Therefore, when n is bounded, and $A_p = \frac{n}{p}\sum_{i=1}^{p}Y_i$, it is known that

$$\frac{A_p-n\mu}{n\sigma/\sqrt{p}} \xrightarrow{\mathcal{D}} N(0,1), \text{ as } p\to\infty.$$

If we can prove that as $p\to\infty$, $\sqrt{p}\,B_p \xrightarrow{\mathcal{P}} 0$ and $\sqrt{p}\,C_p \xrightarrow{\mathcal{P}} 0$, then,

$$\frac{T_p-n\mu}{n\sigma/\sqrt{p}} \xrightarrow{\mathcal{D}} N(0,1), \text{ as } p\to\infty, \tag{I.5}$$

since

$$\begin{aligned} \frac{T_p-n\mu}{n\sigma/\sqrt{p}} &= \frac{(A_p+B_p+C_p)-n\mu}{n\sigma/\sqrt{p}} \\ &= \frac{A_p-n\mu}{n\sigma/\sqrt{p}}+\frac{B_p}{n\sigma/\sqrt{p}}+\frac{C_p}{n\sigma/\sqrt{p}} \end{aligned}$$

with n bounded.

I.2 Proof of $\sqrt{p}\, B_p \xrightarrow{\mathcal{P}} 0$, as $p \to \infty$.

In this section, we prove that $\sqrt{p}\, B_p \xrightarrow{\mathcal{P}} 0$, as $p \to \infty$.

By definition, we have

$$E(B_p) \;=\; \frac{n}{p} E \sum_{i=1}^{p} \int_{-\infty}^{\infty} \gamma_i^2(x) dx. \tag{I.6}$$

Note that, since the integrand is non-negative, the order of integration and expectation in (I.6) can be reversed to give the following form:

$$\begin{aligned}
&= \frac{n}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} E\gamma_i^2(x) dx \\
&= \frac{n}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} \text{Var}\left[\hat{f}_h^{(i)}(x)\right] dx \;\; \text{(by (I.4))} \\
&= n \int_{-\infty}^{\infty} \text{Var}\left[\hat{f}_h^{(1)}(x)\right] dx \\
&= n \int_{-\infty}^{\infty} \frac{1}{n(p-1)} \text{Var}\left[\frac{1}{h} K\left(\frac{x - Z_{11}}{h}\right)\right] dx \\
&= \frac{C}{p-1}, \qquad\qquad (I.7)
\end{aligned}$$

where $C = \int_{-\infty}^{\infty} \text{Var}\left[\frac{1}{h} K\left(\frac{x-Z_{11}}{h}\right)\right] dx < \infty$.

Therefore, we have $\sqrt{p}\, B_p \xrightarrow{\mathcal{P}} 0$, as $p \to \infty$, since

$$\begin{aligned}
\forall \epsilon > 0,\; P(\sqrt{p}\, B_p > \epsilon) &\leq \frac{\sqrt{p}\, E(B_p)}{\epsilon} \\
&= \frac{\sqrt{p}}{\epsilon} \frac{C}{p-1} \;\; \text{(by (I.7))} \\
&\to 0, \;\text{as } p \to \infty.
\end{aligned}$$

Q.E.D.

I.3 Proof of $\sqrt{p}\, C_p \xrightarrow{\mathcal{P}} 0$, as $p \to \infty$.

By definition,

$$C_p = \frac{-2n}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} \delta_i(x)\gamma_i(x)dx.$$

We may also write it as follows,

$$C_p = \frac{1}{p} \sum_{i=1}^{p} I_i,$$

where $I_i \equiv -2n \int_{-\infty}^{\infty} \delta_i(x)\gamma_i(x)dx$.

Thus, $C_p^2 = \frac{1}{p^2}(\sum_{i=1}^{p} I_i)^2$, and

$$\begin{aligned} p\, EC_p^2 &= p\, E\frac{1}{p^2}\left(\sum_{i=1}^{p} I_i\right)^2 = \frac{1}{p}E\left(\sum_{i=1}^{p}\sum_{j=1}^{p} I_i I_j\right) \\ &= \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} E(I_i I_j) \\ &= \frac{1}{p}\left[p\, E(I_1^2) + p(p-1)E(I_1 I_2)\right] \\ &= E(I_1^2) + (p-1)E(I_1 I_2). \end{aligned}$$

I.3.1 Proof of $E(I_1^2) \to 0$, as $p \to \infty$

Note that

$$\begin{aligned} I_1^2 &= 4n^2\left[\int_{-\infty}^{\infty} \delta_1(x)\gamma_1(x)dx\right]^2 \\ &\leq 4n^2 \int_{-\infty}^{\infty} \delta_1^2(x)dx \int_{-\infty}^{\infty} \gamma_1^2(x)dx \end{aligned}$$

by the Cauchy-Schwartz Inequality. Also, we have

$$\begin{aligned} E\int_{-\infty}^{\infty} \delta_1^2(x)dx &= EY_1 = \mu. \\ E\int_{-\infty}^{\infty} \gamma_1^2(x)dx &= \int_{-\infty}^{\infty} E\gamma_1^2(x)dx \\ &= \int_{-\infty}^{\infty} \text{Var}[\hat{f}_h^{(1)}(x)]dx \\ &= \frac{C}{p-1} \to 0. \text{ (by (I.7))} \end{aligned}$$

Therefore, $0 \leq E(I_1^2) \leq 4n^2 \cdot \mu \cdot \frac{C}{p-1} \to 0$, such that $E(I_1^2) \to 0$, as $p \to \infty$.

I.3.2 Proof of $(p-1)E(I_1 I_2) \to 0$*, as* $p \to \infty$

Notice that:

$$\begin{aligned}
\hat{f}_h^{(1)}(x) &= \frac{1}{n(p-1)h} \sum_{i=2}^{p} \sum_{j=1}^{n} K\left(\frac{x - Z_{ij}}{h}\right) \\
&= \frac{1}{n(p-1)h} \sum_{j=1}^{n} K\left(\frac{x - Z_{2j}}{h}\right) + \frac{1}{n(p-1)h} \sum_{i=3}^{p} \sum_{j=1}^{n} K\left(\frac{x - Z_{ij}}{h}\right) \\
&= \frac{1}{p-1} \hat{f}_h(x|2) + \frac{p-2}{p-1} \frac{1}{n(p-2)h} \sum_{i=3}^{p} \sum_{j=1}^{n} K\left(\frac{x - Z_{ij}}{h}\right) \\
&= \frac{1}{p-1} \hat{f}_h(x|2) + \frac{p-2}{p-1} \hat{f}_h^{(1,2)}(x).
\end{aligned}$$

Similarly,

$$\hat{f}_h^{(2)}(x) = \frac{1}{p-1} \hat{f}_h(x|1) + \frac{p-2}{p-1} \hat{f}_h^{(1,2)}(x).$$

Thus,

$$\begin{aligned}
\gamma_1(x) &= f(x;h) - \hat{f}_h^{(1)}(x) = f(x;h) - \frac{1}{p-1} \hat{f}_h(x|2) - \frac{p-2}{p-1} \hat{f}_h^{(1,2)}(x) \\
&= \frac{p-2}{p-1} \left[f(x;h) - \hat{f}_h^{(1,2)}(x) \right] + \frac{1}{p-1} \left[f(x;h) - \hat{f}_h(x|2) \right] \\
&= \frac{p-2}{p-1} \gamma_{1,2}(x) + \frac{1}{p-1} \delta_2(x), \\
\gamma_2(x) &= f(x;h) - \hat{f}_h^{(2)}(x) = \frac{p-2}{p-1} \gamma_{1,2}(x) + \frac{1}{p-1} \delta_1(x).
\end{aligned}$$

Therefore, we compute $E(I_1 I_2)$ as follows:

$$
\begin{aligned}
E(I_1 I_2) &= 4n^2\, E\left[\int_{-\infty}^{\infty} \delta_1(x)\gamma_1(x)dx \int_{-\infty}^{\infty} \delta_2(x)\gamma_2(x)dx\right] \\
&= 4n^2 E\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\gamma_1(x)\delta_2(y)\gamma_2(y)dydx\right] \\
&= 4n^2 E\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)[\frac{p-2}{p-1}\gamma_{1,2}(x) + \frac{1}{p-1}\delta_2(x)] \\
&\quad \cdot\delta_2(y)[\frac{p-2}{p-1}\gamma_{1,2}(y) + \frac{1}{p-1}\delta_1(y)]dydx\} \\
&= 4n^2 E\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)[\frac{(p-2)^2}{(p-1)^2}\gamma_{1,2}(x)\gamma_{1,2}(y) + \frac{p-2}{(p-1)^2}\delta_2(x)\gamma_{1,2}(y) \\
&\quad +\frac{p-2}{(p-1)^2}\delta_1(y)\gamma_{1,2}(x) + \frac{1}{(p-1)^2}\delta_1(y)\delta_2(x)]dydx\} \\
&= 4n^2 E\left[\frac{(p-2)^2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)\gamma_{1,2}(x)\gamma_{1,2}(y)dydx\right] \\
&\quad +4n^2 E\left[\frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)\delta_2(x)\gamma_{1,2}(y)dydx\right] \\
&\quad +4n^2 E\left[\frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)\delta_1(y)\gamma_{1,2}(x)dydx\right] \\
&\quad +4n^2 E\left[\frac{1}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)\delta_1(y)\delta_2(x)dydx\right].
\end{aligned}
$$

Note that

$$
\begin{aligned}
& E\left[\frac{(p-2)^2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta_1(x)\delta_2(y)\gamma_{1,2}(x)\gamma_{1,2}(y)dydx\right] \\
= \;& \frac{(p-2)^2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\left[\delta_1(x)\delta_2(y)\gamma_{1,2}(x)\gamma_{1,2}(y)\right]dydx \\
= \;& \frac{(p-2)^2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[\delta_1(x)]E[\delta_2(y)]E[\gamma_{1,2}(x)\gamma_{1,2}(y)]dydx \\
= \;& 0. \text{ (by (I.2))}
\end{aligned}
$$

Similarly,

$$\begin{aligned} & E\left[\frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta_1(x)\delta_2(y)\delta_2(x)\gamma_{1,2}(y)dydx\right] \\ = & \frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E[\delta_1(x)]E[\delta_2(y)\delta_2(x)]E[\gamma_{1,2}(y)]dydx \\ = & 0. \end{aligned}$$

$$\begin{aligned} & E\left[\frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta_1(x)\delta_2(y)\delta_1(y)\gamma_{1,2}(x)dydx\right] \\ = & \frac{p-2}{(p-1)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E[\delta_1(x)\delta_1(y)]E[\delta_2(y)]E[\gamma_{1,2}(x)]dydx \\ = & 0. \end{aligned}$$

Therefore,

$$\begin{aligned} E(I_1I_2) &= 4n^2\frac{1}{(p-1)^2}E\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta_1(x)\delta_2(y)\delta_1(y)\delta_2(x)dydx\right] \\ &= \frac{4n^2}{(p-1)^2}E[\int_{-\infty}^{\infty}\delta_1(x)\delta_2(x)dx]^2 \\ &\leq \frac{4n^2}{(p-1)^2}E\left[\int_{-\infty}^{\infty}\delta_1^2(x)dx\int_{-\infty}^{\infty}\delta_2^2(x)dx\right] \text{ (by the Cauchy Schwartz Inequality)} \\ &= \frac{4n^2}{(p-1)^2}E\left[\int_{-\infty}^{\infty}\delta_1^2(x)dx\right]E\left[\int_{-\infty}^{\infty}\delta_2^2(x)dx\right] \\ &= \frac{4n^2}{(p-1)^2}\int_{-\infty}^{\infty}E\delta_1^2(x)dx\int_{-\infty}^{\infty}E\delta_2^2(x)dx \\ &= \frac{4n^2}{(p-1)^2}\int_{-\infty}^{\infty}\text{Var}[\hat{f}_h(x|1)]dx\int_{-\infty}^{\infty}\text{Var}[\hat{f}_h(x|2)]dx \text{ (by (I.3))} \\ &= \frac{4n^2}{(p-1)^2}\int_{-\infty}^{\infty}\frac{1}{n}\text{Var}\left[\frac{1}{h}K\left(\frac{x-Z_{11}}{h}\right)\right]dx\int_{-\infty}^{\infty}\frac{1}{n}\text{Var}\left[\frac{1}{h}K\left(\frac{x-Z_{21}}{h}\right)\right]dx \\ &= \frac{4}{(p-1)^2}C_1C_2, \end{aligned}$$

where

$$\begin{aligned} C_1 &= \int_{-\infty}^{\infty}\text{Var}\left[\frac{1}{h}K\left(\frac{x-Z_{11}}{h}\right)\right]dx < \infty, \\ C_2 &= \int_{-\infty}^{\infty}\text{Var}\left[\frac{1}{h}K\left(\frac{x-Z_{21}}{h}\right)\right]dx < \infty. \end{aligned}$$

Therefore,

$$(p-1)E(I_1I_2) = (p-1)\frac{4}{(p-1)^2}C_1C_2 = \frac{4}{p-1}C_1C_2 \to 0,$$

as $p \to \infty$.

I.3.3 Proof of $p\ EC_p^2 \to 0$, as $p \to \infty$.

By combining results in section I.3.1 and I.3.2, the proof is obvious.

It follows that

$\forall \epsilon > 0$,

$$\begin{aligned} P(\sqrt{p}\, C_p > \epsilon) &= P(p\, C_p^2 > \epsilon^2) \\ &\leq \frac{p\ EC_p^2}{\epsilon^2} \\ &\to 0, \end{aligned}$$

as $p \to \infty$, by (I.3.3).

Q.E.D.

APPENDIX II

SOMETHING ABOUT CONVOLUTION

We used the property of convolution of two normal distributions for the estimation of μ. In this appendix, we introduce something about the convolution.

The convolution $f * g$ of two functions $f(x)$ and $g(x)$ defined in R is given by:

$$f * g(z) = \int_R f(x)g(z-x)dx.$$

In probability theory, the convolution of two functions has a special relationship with the distribution of the sum of two independent random variables. If the two random variables X and Y are independent, with pdf's f and g respectively, the distribution $h(z)$ of $Z = X + Y$ is given by $h(z) = f * g$. This result is obtained below.

$$\begin{aligned} H(z) &= P(Z \leq z) = P(X + Y \leq z) \\ &= \int P(X + Y \leq z | Y = y) \cdot g(y)dy \\ &= \int P(X \leq z - y) \cdot g(y)dy \\ &= \int F_X(z - y) \cdot g(y)dy \end{aligned}$$

$$\begin{aligned} h(z) &= \frac{H(z)}{dz} = \frac{d(\int F_X(z-y) \cdot g(y)dy)}{dz} \\ &= \int \frac{d(F_X(z-y))}{dz} \cdot g(y)dy \\ &= \int f(z-y) \cdot g(y)dy \\ &= f * g \end{aligned}$$

Given two normal probability density functions with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively, the convolution of these two normal functions is also a normal probability density function with mean $(\mu_1 + \mu_2)$, and variance $(\sigma_1^2 + \sigma_2^2)$.

Denote $G_1 \equiv \phi(x; \mu_1, \sigma_1^2)$ and $G_2 \equiv \phi(y; \mu_2, \sigma_2^2)$. The convolution of G_1 and G_2 is:

$$G_1 * G_2(z) = \phi(z; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

APPENDIX III

R FUNCTION **TS** FOR CALCULATING TEST STATISTIC $T^{(S)}$

III.4 Description

The following R function **TS** computes the test statistic $T^{(S)}$.

1. z is the data matrix that should have p rows and n columns.
2. x, y are both vectors.
3. h is the over-smoothed bandwidth.
4. The outputs are: unstandardized test statistic: $T_p^{(S)}$; σ_S ; standardized test statistic: $T^{(S)}$.

III.5 R Functions Used in TS

```
h1=function(x,h){
n=length(x)
X=matrix(1,n,1)
X=X-x
X=dnorm(X(sqrt(2)*h))
stat=sum(X)-n*dnorm(0)
stat(sqrt(2)*h*n*(n-1))
}
h2=function(x,y,h){
n=length(x)
X=matrix(1,n,1)
```

```
X=X-y
X=dnorm(X(sqrt(2)*h))
stat=sum(X)
stat(sqrt(2)*h*nˆ 2)
}
rphat=function(Z,i,h){
p=nrow(Z)
n=ncol(Z)
vec=(1:p)[(1:p)!=i]
rphat=0
for(j in vec){
rphat=rphat+h2(Z[i,],Z[j,],h)
}
rphat(p-1)
}
```

III.6 R Function TS

```
TS=function(Z,h){
p=nrow(Z)
n=ncol(Z)
H1=1:p
Rphat=1:p
for(i in 1:p){
H1[i]=h1(Z[i,],h)
Rphat[i]=rphat(Z,i,h)
```

```
}
S.W=mean(H1)
S.B=mean(Rphat)
Tp=S.W-S.B
data=H1-2*Rphat
Shat=sd(data)
print(c(Tp,Shat,sqrt(p)*TpShat))
}
```

VITA

Dongling Zhan was born in Shanghai, P. R. China. She received her B.S. degree in Mathematics Statistics from Fudan University, in Shanghai, P. R. China. She was a volunteer teaching English in Western China before entering the master's program in the Department of Statistics at Texas A&M University, College Station, Texas. After she graduated with a M.S. degree, she continued pursuing a Ph.D. degree in Statistics at Texas A&M University, which was completed under the advisement of Professor Jeffrey D. Hart in May 2012.

Her mailing address is:

Department of Statistics

Texas A&M University

3143 TAMU

College Station, TX 77843-3141

C/O Jeffrey D. Hart, Ph.D.